Volume 58
Number 2

Whole No. 267
1945

Psychological Monographs

JOHN F. DASHIELL, *Editor*

Improvability of Pitch Discrimination

By
RUTH F. WYATT, Ph.D.
Northwestern University

Published by
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Publications Office
NORTHWESTERN UNIVERSITY, EVANSTON, ILLINOIS

ACKNOWLEDGEMENTS

THE WRITER wishes to acknowledge with sincere gratitude the valuable guidance of Professor R. H. Seashore throughout this experiment. Grateful acknowledgement is also made to Professor W. L. Valentine for helpful counsel in connection with the preparation of the manuscript and to Dean J. W. Beattie, who assisted in the selection of subjects.

TABLE OF CONTENTS

TABLE OF CONTENTS

I. [illegible]

II. Review of Research

Introduction

Effects of [illegible]

Effects of [illegible] Musical [illegible]

Effects of [illegible] Reading

[illegible]

III. [illegible] of the Present Study

[illegible]

[illegible]

[illegible]

IV. [illegible]

[illegible]

[illegible]

V. [illegible]

[illegible]

VI. [illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

IMPROVABILITY OF PITCH DISCRIMINATION

I. Theoretical Considerations

THE *"capacity" hypothesis and the "physiological limit."* It is held by many psychologists that our powers of sensory discrimination have fixed limits which depend upon inherited structural features of the receptors. In the field of audition, for example, individual differences in the power to discriminate pitch, *i.e.,* to discern differences or to identify the direction of differences in pitch, have been ascribed to inherited structural characteristics of the auditory sensorium. According to this view, each individual has a maximum potential power or "capacity" for such discrimination and this capacity, being determined by the inherited efficiency of the ear, is regarded as "set" early in childhood and unsusceptible of improvement through environmental influences.[1]

Closely related to this definition of a capacity is the concept of a "physiological limit." The *measurement* of any capacity necessarily involves its quantitative expression and in pitch discrimination, the "physiological limit" has been expressed as a psychophysical threshold, i.e., in cycles per second, or as a score or rank in a test. Whatever the form of designation, the physiological limit represents the quantification of the bed-rock limit of an individual's capacity.

In the field of psychology of music and particularly in connection with pitch discrimination, the concepts of a capacity and of a physiological limit have been brought into special prominence by Seashore.[2] In 1919, he wrote (24, p. 57):

> Pitch discrimination is not a matter of logical judgment. It is rather an immediate impression, far more primitive than reflective thought, and dependent upon the presence or absence in various degrees of the sensitive mechanism in the inner ear.

In 1938, Seashore defined the physiological limit as "that limit for sensation and perception which is set by the structure of the sense organ and the brain" (26, p. 57) and stated his view that although it may vary within a small range with such factors as "fatigue, rest, the action of either depressive or stimulative drugs, or disease" (p. 59), the physiological limit for pitch discrimination does not vary with age or training.

> It seems probable that just as the physical eye of the child at the age of three is as keen as it ever will be, so the pitch sensitiveness in the ear probably reaches its maximum very early. Development in the use of the sense of pitch with maturation consists in acquiring habits and meanings, interests, desires, and musical knowledge, rather than in the improvement of the sense organ. . . . The physiological limit for hearing pitch does not improve with training. Training, like maturation, results in the conscious recognition of the nature of pitch, its mean-

[1] The terms "capacity" and "ability" have had distinct meanings in psychology. A "capacity" is defined as "potentiality of the organism as provided and limited by native constitution" (37, p. 1). The term "ability" is conventionally used to designate acquired skill in the *use* of a capacity (24, p. 15). Thus, unlike capacities, abilities are believed to be susceptible of improvement.

[2] C. E. Seashore is cited as the foremost authority in connection with matters bearing on the psychology of music and the measurement of musical talent. Early in the century he formulated a program of musical guidance which is still in use in many localities. In his long and distinguished career at the University of Iowa, Seashore has conducted and directed voluminous research in the field of psychology of music and particularly on the measurement of musical talent. Since he stands as a pioneer in these fields, and since his systematic position has been so influential, Seashore's views are taken as a point of reference.

ing, and the development of habits of use in musical operations. Training probably does not modify the capacity of the sense organ any more than the playing of the good violin may improve the quality of its tone (p. 58).

In 1940 (28, p. 36) Seashore re-defined the physiological limit as:

> that limit of achievement which is set by characteristics of the organism and beyond which training is relatively ineffective. This is, however, not a fixed limit because organic changes take place from the beginnings of embryonic life throughout the period of maturation. It may, however, be regarded as representing that capacity for which individual differences are largely dependent upon heredity.

The measurement of pitch discrimination and the concept of the "cognitive limit." In their construction and also in their intended applications, the Seashore Measures of Musical Talents[3] follow the thesis that the physical characteristics of the sound wave—frequency, amplitude, form and duration—are the only variables by which the performer can convey music *per se* to the listener and that therefore the psychological correlates of these physical variables, viz., pitch, loudness, timbre and time, are of fundamental importance in the appraisal of musical talent. The tests reflect their author's desire to divest them of all musical "meanings" in order to make them measures of natural capacity for musical growth regardless of the amount of previous musical experience. In the tests of pitch discrimination, for example, an attempt is made to hold constant the intensity, the timbre and the time relationships. The tests are thus reduced to their simplest form, *Ss* being required merely to state whether the second of two successive pure tones is higher or lower than the first. The tests were made "elemental" in order to measure as closely as possible the "physiological limit" of a "capacity." Due to unfavorable factors in the measurement process, however, it is conceded that they may not actually succeed in disclosing this ultimate limit in all cases.

Seashore has repeatedly cautioned that although the aim of the tester should always be the determination of an individual's maximum potential capacity, the result may actually fall short of this ideal, producing what he has termed a "cognitive limit." A review of Seashore's publications up to 1940 indicates that failures to ascertain the physiological limit were ascribed primarily to the intrusion of unfavorable factors which were related to *cognition* rather than *audition*. In 1910, Seashore defined the cognitive limit as

> . . . a higher threshold [higher than the physiological threshold] due to lack of information, best form of attention, interest, effort, etc.; or to disturbances of some sort (17, p. 49).

In 1919, it was defined as

> . . . an inferior record due to some difficulties or disturbances, such as distraction, ignorance, and lack of power of application (24, p. 51).

and in 1938 (26, p. 57), Seashore still maintained that the cognitive limit

> . . . usually is due to a lack of understanding of the test requirements, or a lack of mental development, or of good will, or of general power of application on the part of the subject tested.

In the concluding section of the 1940

---

[3] The content of the original battery of tests which appeared in 1919 is well known. The 1939 revision (27, 28) has certain formal differences, but in their content, method and proposed applications, the revised measures are governed by substantially the same psychological assumptions as those which formed the groundwork for their predecessors. The revised battery contains tests for tonal memory and for pitch, loudness, timbre, time and rhythm discrimination.

monograph (28), Seashore's views appear to be somewhat modified. A distinction is still made between the physiological and the cognitive limits, but the latter is somewhat more broadly conceived than in the earlier publications. The nature of the cognitive factors is not specified in detail, but in addition to factors which appear to be similar to the conventional ones mentioned in earlier papers, several references to "work method"[4] may be found (p. 40):

> Achievement in each and all of the measures is subject to improvement with training insofar as insight into the testing situation, ability to comprehend the task of learning, ability to concentrate on the specific issue in listening, and favorable environmental conditions are concerned. Rhythm and memory are more subject to such improvement than the four more elemental measures. . . .
>
> When changes in rating are analyzed, the improvement with practice is often traceable to change in work method, not in actual spontaneous hearing. Even in such simple tasks as those here involved, the method of listening and making judgments may vary in many respects, some being better or worse than others. Choice of the better work method may show rise in the practice curve; but such change in achievement does not necessarily indicate any change in capacity for hearing. The work method is often influenced by attitude, division of labor, tendency to anticipate, and lazy or indifferent resort to guessing.

The cognitive factors have not been regarded as necessary evils. In 1938, Seashore stated (26, p. 57) that the margin between the physiological and the cognitive limits

> may be reduced or eliminated by a repetition [of the test] and by individual testing by an expert.

and that

> . . . A good test in the hands of an expert may properly establish the physiological limit of pitch discrimination in the first trial for a majority of the subjects in a group test, whereas in an individual test the physiological limit may be determined with a high degree of certainty for practically all.

By 1940, however, Seashore cautioned that when individuals with poor or doutful records are retested,

> . . . there should be intensive preliminary practice to serve as a means of diagnosing difficulties encountered and this should be extended in proportion to the seriousness of the difficulties. Without such diagnosis, retesting in the case of a poor record may be futile (28, p. 39).

It is apparent from the above that the term "cognitive limit" has not had a static or standardized meaning. Though interpreted for three decades to include principally factors which would affect scores in *any* test (i.e., distraction, lack of application, poor motivation, failure to understand instructions), in the 1940 monograph the attainment of a cognitive, rather than a physiological, limit seems to be attributed to almost any factor, subjective or objective, which might inhibit a threshold, score or rank representing bed-rock capacity.[5]

This new concept of the cognitive limit does not appear to be clearly documented nor, in view of the rather significant extension of the earlier point of view, does it seem adequately elaborated. It is difficult, therefore, to give concrete content to such expressions as "ability to comprehend the task of learning" or "division of labor" in their relation to the concept of the cognitive limit.

Other views. Not all psychologists concur in these views. Pratt (16) insists that no psychological experience can be iden-

[4] *Cf.* p. 4.

[5] These elements seem so varied as to make questionable the appropriateness of the term "cognitive."

tified with the conditions enumerated in its explanation. "Most of the conditions of experience," he says (p. 61) "exist in the physiological processes of the organism and in the physical aspects of the stimulus, but in neither of these domains are the psychological properties of experience to be found." Similarly, Mursell (13), rejecting what he calls Seashore's sensationalistic position, contends that the experience of pitch does not depend simply upon the action of the ear, but involves the "integrating, selecting, interpreting action of the central nervous system" (p. 76). He denies that the apprehension of pitch can be explained merely by reference to the frequency of the sound wave and the action of the inner ear and regards as an "unproved assumption" the claim that pitch discrimination cannot be improved by training.

R. H. Seashore (30) has suggested a modified concept of the physiological limit. He does not conclude that there are no physiological limits, but maintains that such limits may be valid only for a given work method and that if better work methods are used, significant improvement may be obtained. Work methods are defined as "patterns of behavior which are adopted by each individual in the course of utilizing his biological equipment during learning" (p. 123). They include "any variation in set, attitude, approach, trick of the trade, adjustment mechanisms, etc., in other words qualitative variations in *ways* of reacting to a situation" (p. 124).[6] R. H. Seashore has emphasized that such work methods may frequently be adopted without *S*'s awareness of their nature.

According to this work methods hypothesis, even after ordinary precautions are taken with respect to control of the testing conditions and ensuring of *S*'s understanding of the instructions, adoption of better work methods may lead to further improvement:

> In measuring individual differences it is not sufficient to control the instructions or working situation, for the observer's previous incidental background may lead him to adopt very different work methods from those expected. It follows that 'control' limited to ordinary instructions and demonstrations is incomplete, and that other unnoticed factors operate to modify the work method actually adopted (p. 123).

Importance of pitch discrimination. Pitch discrimination figures prominently in most discussions of musical talent. Seashore in particular has always stressed its importance and its validity for prognostic purposes. He regards his pitch test as basic, not only to the hearing of tones, melodies, harmonies and timbres, but also to good musical performance, to musical imagery and memory and even to music appreciation. All of these "derived factors" presumably depend upon the capacity for pitch discrimination. In 1919 Seashore wrote (24, p. 42):

> [Pitch discrimination] is a fundamental capacity in musical talent, and upon it rest most of the powers of appreciation and expression in music. One must hear pitch

[6] In a very recent article not included in the bibliography, the term "capacity" has been defined so as to include these modifications suggested by the "work methods" hypothesis:

"A person's functional capacity for a given performance or skill is his maximal, *potential* effectiveness in terms of end results. This must be conceived, however, on the basis of a given work method, with the understanding that if the work method is changed a different capacity may be called into play. It is also assumed that a person will be able to perform at capacity (for a given developmental level) only after he has had *optimal training under optimal motivation.* Since capacity refers to a *potential* limit, it can only be *inferred*, and hence not operationally defined in such a way as to be readily observable." (from Jones, H. E. and Seashore, R. H. The development of fine motor and mechanical abilities. *43d Yearb. Nat. Soc. Stud. Educ.*, 1944, Part I, p. 134.

differences in order to appreciate tones. One must be guided by such hearing in playing and singing. The imagining, the remembering, the thinking about, and the arousal of feeling for tones are all limited by the capacity for hearing differences of pitch.

In a recent monograph (28, p. 46) he has reaffirmed these views and made them still more specific, declaring

. . . that successful musicians almost without exception reveal a fine sense of pitch; that a good or a poor sense of pitch at the beginning of musical education significantly predicts correspondingly good or poor progress in the mastery of pitch; that persons with a fine sense of pitch are correspondingly critical in the judgment of pitch performance; that there is a tendency for persons with a fine sense of pitch to succeed with musical instruments which demand it; and that numerous cases show that an unsatisfactory sense of pitch accounts for musical failure and discouragement.

Implications for guidance and pedagogy. The present problem has important practical implications for guidance and for musical education. The reasoning which underlies the use of talent tests in connection with guidance may be summarized as follows: (1) tests have been constructed which purport to constitute a valid and basic measure of certain musical talents; (2) pitch discrimination is regarded as the most important and basic of all such measures; (3) if a good pitch discrimination test is properly administered under ideal conditions, it is possible to determine the "physiological limit" or the "approximate physiological limit" of the "capacity" for pitch discrimination; (4) by definition, the physiological limit cannot be improved through training; (5) a score or rank in such a test therefore constitutes a legitimate basis for guidance, either for encouragement of those who seem gifted or for the preclusion of failure and disappointment on the part of those who are apparently ungifted.

The validity of these propositions is of considerable importance in music education, for if training is indeed futile, emphasis should be placed upon guidance. This is Seashore's stand. He refutes the view that music lessons are a remedy for all and deplores the failure to recognize the limits of educability imposed by what he believes are relatively fixed individual differences. As early as 1910, he wrote (17, p. 58):

What a blessing to a girl of the age of eight, if the music teacher would examine her, and, if necessary say, "much as I regret it, I must say that you would find music dull and difficult, and I would advise you to take up some other art." What a blessing if that child could be started right; but current theory and practice is against her. There is too much faith in what music lessons can do for a person without native capacity. If we are to have musical ears, we must be born with them. That is the probable finding of current research.

He has recently declared with even greater emphasis (26, p. 58):

Fortunes have been spent and thousands of young lives have been made wretched by application of the theory that the sense of pitch can be improved with training. It is the cause of the outstanding tragedy in musical education.

The other side of the question was stated as early as 1903 by Whipple (38, pp. 303-304):

I believe that it is still an open question, and one worthy of solution, as to whether musical incapacity, especially when discovered in early childhood, may not be remedied by proper training. . . . A new incentive . . . would be given the musical training of children if we knew that environmental, rather than hereditary or innate, influences were responsible for the closing of one of the great avenues of aesthetic expression and enjoyment.

If crucial experimental evidence were to indicate that after competent testing no form of practice or training resulted in significant improvement, then prognosis on the basis of the initial test or tests would seem fully justified. On the other hand, if the evidence were to show that training frequently resulted in such significant improvement that prognosis based on the original performance would have been misleading, then attention had best be focussed upon early diagnosis of individual difficulties and discovery of the best remedial methods for overcoming them. Under these conditions the tests would not lose their practical usefulness, for they would still make an important contribution when directed toward diagnostic ends.

II. Review of Research

INTRODUCTION

The experimental literature bearing on the question of improvability of pitch discrimination does not yield a clear-cut answer to the problem. A number of published investigations on the effects of training present data and conclusions which appear to support Seashore's "capacity" hypothesis. On the other hand, some experiments are available in which the evidence and interpretations are in conflict with Seashore's position.

A large part of this inconclusiveness may be attributed to the great diversity of training procedures. Farnsworth (5, p. 245) aptly observes that "data on the effects of training will have different meanings as the meaning of training varies." The experimental literature will here be classified into the following three groups according to the type of training given:

(1) Effects of practice
(2) Effects of formal musical instruction
(3) Effects of remedial training.

While these methods have not been and need not be mutually exclusive, it is helpful to consider them separately in reviewing and appraising their contribution to the problem.

In the studies selected for inclusion under the first heading, the technique has consisted merely in repeating the tests as many as twenty times with little or no variation in method, *E* noting the progressive scores or thresholds during the practice series or comparing initial and final results. The second group includes studies in which *Ss* were tested before and after instruction in applied and theoretical music. Studies subsumed under the heading of remedial training are more varied in nature and include such techniques as the following: (1) informing *Ss* as to the correctness of their responses; (2) demonstrations with foreknowledge of the correct answer; (3) drill in discrimination or recognition of piano intervals; (4) vocal matching of tones, sequences or scales; (5) other techniques of illustration, explanation or suggestion. In most of the studies of this type, *Ss* were given individual help and an effort was often made to adapt the training to their particular difficulties.

EFFECTS OF PRACTICE

A study of the effects of continued repetition of a test without resort to remedial instruction does not at first glance appear relevant to an investigation of the effects of training. Nevertheless, the failure of *Ss* to improve with repeated administrations of a test has actually been proferred as evidence that the physiological limit of pitch discrimination is not susceptible of improvement through training (24, p. 60).

H. S. Buffum. An unpublished experiment performed sometime prior to 1910 by Buffum is described by Seashore (17, 24) and by Smith (31). Using tuning forks struck on a sounder and held to a resonator, Buffum first gave 15-minute individual tests to 25 eighth grade children. After determining their thresholds, presumably at a standard frequency level of 435 ~, the *Ss* were classified into three groups with modes at 3 ~, 8 ~, and 17 ~ respectively. They were then given twenty 40-minute periods of "specific and intensive practice" (24, p. 60), *E* taking records by the method of right and wrong cases (17, p. 54). It was found that 23 of the 25 children remained in their original

classification. Of the remaining two, one was a borderline case and the other, presumably because of misunderstanding the preliminary test, was placed originally in the lowest group but changed during the practice to the highest.

Seashore regarded this experiment as evidence of the nonimprovability of pitch discrimination, pointing out that 24 of the 25 cases had evidently revealed their physiological thresholds in the first test, since practice thereafter was apparently of no avail. But alternative interpretations might be that even with as many as 20 periods of identical repetition of a test, (1) *Ss* may have remained on what Smith terms a "cognitive plateau" or, (2) along the lines suggested by R. H. Seashore's work methods hypothesis (30), the failure to improve could be construed as evidence of the stability of pitch thresholds under constant (and possibly defective) work methods.

F. O. Smith. An early (1914) experiment by Smith (31) purports to investigate, among other matters, the effects of practice upon pitch discrimination. Its significance is somewhat beclouded, however, by a failure on the part of the author to adhere to clear differentiations in terminology and by an imperfect segregation of data. In 1914 the terms "practice," "instruction" and "training" were often used interchangeably and Smith, like many others writing at that time, was not always careful to separate the results according to the "training" technique employed.

That portion of Smith's paper which deals with the effect of practice describes an experiment on 476 children who were given a series of 12 pitch discrimination tests. The experiment is complicated by the fact that the lowest 106 cases were also given supplementary individual assistance (instruction) concurrently with the practice series. Average thresholds for these 106 "instructed" *Ss* are presented separately in Smith's report,[7] but the results for the remaining 370 *Ss* are not separately given. Accordingly, this portion of Smith's experiment cannot be considered as a characteristic example of the effects of practice alone.

The tests were given with a standard tuning fork of 435 ~ and ten incremental forks respectively 30, 23, 17, 12, 8, 5, 3, 2, 1 and 0.5 ~ higher than the standard. The forks were struck on a sounder and held to a resonator. Each of the 12 tests was preceded by a brief warming-up exercise in which *Ss* answered orally. The average threshold for the 215 boys dropped from 8.1 ~ to 4.7 ~, while that for the 261 girls dropped from 6.3 ~ to 4.5 ~. Of the 476 children, 270 (57%) improved considerably, reducing their average threshold from 7.5 ~ to 2.9 ~.[8] There is no indication in Smith's study as to how many of the 106 "instructed" *Ss* belong in the group of 270, but even if they are *all* subtracted, there would still remain 164 *Ss*–34% of the original group of 476–who improved with practice alone.

A table based on "internal evidences" shows "the distribution of those who reach the approximate physiological threshold on different days of practice." Only 47% of Smith's *Ss* who improved reached this "limit" by the fifth day and of the remaining 53%, 28% were spread over the last three days (pp. 79-81).

Smith concurs with Seashore in his view that the "sensitiveness of the ear to pitch differences [i.e., the 'physiological

[7] *Vide* p. 18, this paper.

[8] The average thresholds for the twelve tests were 7.5, 7.1, 5.7, 5.0, 5.0, 4.3, 4.1, 3.4, 3.5, 3.0, 3.0 and 2.9 ~ respectively.

limit for pitch discrimination'] can not be improved appreciably by practice" (p. 101). He believes, however, that "instruction in regard to the nature of the test and individual help are all important for the lowering of the cognitive limit," while "mere practice" toward this end is admitted to be "a poor and uncertain makeshift" (*ibid*).

Frances A. Wright. Using the 1919 form of the recorded Seashore pitch test,[9] Wright (41, 42) tested 24 adult music majors for one hour a day for five consecutive days. The students were not given coaching or remedial training of any kind during this period and the papers were not scored until the end of the testing program.

The daily score averages for each of the 24 *Ss* are shown in Table 1, which is reproduced from the original data sheet.[10] For the group as a whole, the mean scores were 85.71, 85.35, 87.02, 87.96 and 87.90 in the five days of testing. The gain seems relatively insignificant and it might be supposed that the first day's testing had disclosed the "approximate physiological limit" for these *Ss*.

An interesting divergence is found, however, when the score fluctuations of the six highest *Ss* in the distribution are compared with those of the six lowest *Ss*.

TABLE 1

Raw score averages for Frances A. Wright's 24 *Ss* in five consecutive days of testing

	Raw Score Averages				
S	1st day	2nd day	3rd day	4th day	5th day
1	81	84.5	84.5	87	86.5
2	80	80.5	85	83.5	91.5
3	87	84.5	90	87.5	86.5
4	87	83.5	86.5	87.5	89.5
5	88	88	88.5		90.5
6	80	88	90.5	86.5	84.5
7	90	85	85.5	86.5	88
8	84	85.5	89	88.5	79
9	90	87.5	89.5	87	88
10	89	88	89	92	
11	82.5	83.5	87	84	91.5
12	87	86	88	98	95.5
13	91	93	84.5	88.5	89
14	88	84	82	92	89.5
15	87	85	83.5	90	
16	85	87	88.5	85.5	87.5
17	88	87	89.5	90.5	90.5
18	85	83.5	85.5	83	87.5
19	89.5	85	89	90.5	86
20	83	84	86	86	
21	90	88	90	90	91
22	83	87	89	90.5	88
23	82	82	82	87.5	81.5
24	80	78.5	86	81	84.5

The daily averages for these two groups are found to be as follows:

	High Group	*Low Group*
1st day	89.92	80.92
2nd day	87.75	82.83
3rd day	87.92	85.83
4th day	89.03	84.91
5th day	88.40	86.67

There is no resemblance to a learning curve in the data for the high group. Their scores may be said merely to fluctuate within a two-point range. No such stability is found for the low group, however. With the exception of the fourth day, their mean scores show a small daily increase. By the fifth day the average gain is almost six points.

In terms of centile rank (23), the six highest *Ss* began with an average rank of 94.8 and dropped to 91.0 by the fifth day. This loss would not be sufficiently significant to alter prognosis. The six lowest *Ss*, on the other hand, began with

[9] The 1919 form of the test was recorded on a double-faced 12-inch phonograph record (23). All of the tonal stimuli were produced by tuning forks. There were 100 trials, in each of which a standard tone of 435 ~ was either the first or the second tone of a pair. The comparison tones were higher by 30, 23, 17, 12, 8, 5, 3, 2, 1 and 0.5 ~ respectively. Two tones were sounded in succession and *Ss* were asked to tell whether the second tone of each pair was higher or lower than the first. Ten trials were given at each of the above increments and the score was the percentage of correct responses. Centile ranks were available for adults, eighth grade and fifth grade children.

[10] The writer wishes to express her appreciation to Miss Wright for having supplied these unpublished data.

a mean centile rank of 49.6 but rose by the last day of practice to a rank of 82.5 —a change by degrees from the mean of the adult population to a rank which Seashore would now classify as "excellent" (27).

Discussion. It may be helpful at this point to summarize the inferences which may be drawn from a review of these studies of the effects of practice and to relate such inferences to the theoretical considerations set forth in the first part of this paper.

In each of the three experiments described under this heading, there are instances of improvement as well as failure to improve. As has been indicated above, there has been a tendency to assume, when *Ss* did not improve, or did not improve greatly, with practice that their "physiological" or "approximate physiological" limits had been measured in the original testing. Several alternative explanations for the absence of improvement might be suggested, however: (1) *Ss* may have failed to improve because they remained on a "cognitive plateau," i.e., in the retesting, such conventional "cognitive" difficulties as low motivation, distraction, failures of understanding, etc., may not have been overcome or sufficiently reduced to change the record materially; (2) *Ss* may have applied a constant or equally inefficient "work method," not necessarily the best, and they may have failed to "hit upon" better *ways* of reacting, e.g., learning to retain a clearer image of the first tone for comparison with the second, employing a visual image of a vertical scale or attempting to reproduce the tones by implicit throat action (30, p. 129); (3) failure of *Ss* to show improvement might be ascribed to the test itself, e.g., *Ss* may be so close to the test's ceiling at the outset, that any potential improvement, ascertainable only with more finely graded or more difficult increments, could not be manifested because of the limitations of the test actually employed; (4) improvement may have been concealed by the method of handling the data. In Buffum's experiment, for example, *Ss* were classified into three groups with modes at 3, 8 and 17 ~. These are relatively coarse groupings and improvement may have occurred without its being sufficient to change the classification. A more common error may be found when data are presented solely in terms of group means. This is most clearly illustrated in the re-working of Wright's data, shown on p. 9. If these findings are characteristic of the effects of repeated testing, it would seem probable that the less proficient *Ss* profit more from retesting than is indicated by the very small increases in group averages. It is possible that intensive retesting might grow wearisome to the initially high *Ss* to about the same extent that it improved the records of the initially low *Ss*.[11] But if the only data given are the means for the entire group, pitch discrimination may appear to be more stable than it actually is.

Turning to instances of improvement with practice, several explanations are possible: (1) In the absence of coaching or instruction, *Ss* may overcome or reduce their own difficulties in "cognition." Smith's data show, however, that "mere practice" improvement was far from immediate for most of the *Ss*. (2) Even when left to their own devices, *Ss* may have employed better work methods. (3) The presence of uncontrolled sources of error in the testing situation must also be considered as a possible factor in ac-

[11] A general tendency for this to occur in a single retest might account, at least in part, for the low retest reliabilities for tests of this character.

counting for improvement with practice. Variation in the intensity or the duration of the tones, peculiarities in the manipulation of tuning forks, localization, *E*'s manner or facial expression, improvement in the physical conditions of the room—any or all of these objective factors might result in changes in performance with successive retests. (4) When improvement in a pitch discrimination test occurred with practice, it was attributed by Seashore to the overcoming or reduction of factors of cognition. He rejected the possibility of physiological or neurological changes except insofar as drugs, disease, fatigue, etc., might have an adverse effect. However, the absence of favorable physiological or neurological change does not appear to have been conclusively demonstrated by the experimental evidence reviewed here. It is suggested, therefore, that pending further evidence, the occurrence of physiological or neurological changes be retained as at least one of the possible theoretical explanations of improvement in pitch discrimination.

In general, the experimental evidence from studies on the effects of practice is inconclusive. But even if the findings were consistently negative, such findings would not by themselves constitute unconditional proof of the unimprovability of pitch discrimination. Judgment must be withheld pending a study of the effects of other types of training. Whipple's comment in an editorial note to an article written by Farnsworth in 1928 (5, p. 240) still seems pertinent:

> The interpretation of results in such mass experiments as those of Buffum and Smith is, in my judgment, decidedly difficult, if not often misleading. Certainly, an equally important method of studying the effects of practice is to confine one's effort to drilling a competent, though unmusical, adult under laboratory conditions which permit some measure of qualitative analysis of what takes place.

EFFECTS OF FORMAL MUSICAL INSTRUCTION

Hazel M. Stanton and Wilhelmine Koerth. The work of these investigators (33, 34, 35) is the most intensive and extended research available on the effects of formal musical instruction on Seashore test scores. These studies have been accepted and sponsored by Seashore as "final critical proof" that his tests measure "capacities," i.e., maximum physiologically-determined potentialities which are not improved by training. In his preface to one of these studies (34), Seashore writes:

> The Measures of Musical Talent here discussed were built on the assumption that they should measure these specific capacities before musical education was begun and that the capacities would not be greatly modified by training. Attempts to validate this assumption have been made by various indirect methods, but the final critical proof has awaited the accumulation of successive measurement upon the same individuals before, during, and after musical training.

The *S*s in these investigations were all enrolled in the Eastman School of Music, either in the Preparatory Department or as special students or music degree majors at the college level. *S*s were divided into the following groups:

(1) 285 pre-adolescents (grades 4, 5, and 6) enrolled in the Preparatory Department
(2) 208 adolescents (grades 7, 8 and 9) enrolled in the Preparatory Department
(3) 152 post-adolescents (special students or enrolled in the Preparatory Department)
(4) 157 music degree majors.

The method of the investigation consisted of administering the Seashore tests to *Ss* at the time of their entrance to the Eastman School and administering retests to the same *Ss* after a three-year interim of musical instruction. This musical instruction varied in amount and character for the different groups. For *Ss* in the Preparatory Department, musical training consisted of two weekly half-hour lessons, one in applied music (piano, violin, clarinet, etc.) and one in "musicianship" (music theory). The music degree majors had three types of musical training: (1) individual lessons in voice or instrument; (2) group training in instrumental and vocal ensemble; (3) general courses typical of any music curriculum at the college level, e.g., conducting, harmony, theory, form, orchestration, counterpoint, history of music.

TABLE 2
Raw scores in two pitch discrimination tests with three years of intervening musical instruction (from Stanton and Koerth)

Ss	Mean T1	Mean T2	Net Change
Gr. 1	76.5	81.2	+4.7
Gr. 2	81.2	83.3	+2.1
Gr. 3	80.6	81.9	+1.3
Gr. 4	84.0	84.1	+0.1

It was assumed by the authors that . . . "If the scores of these students vary little upon retesting, there is not only proof from the practical situation that the Measures will yield stable results, but there is added information regarding the innateness of musical talent and of the capacity nature of these tests" (35, p. 29).

Table 2 shows the mean raw scores in the first (T1) and second (T2) administrations of the Seashore pitch test (1919 form). Three years of musical instruction, as described above, intervened between the two tests. Table 3 presents an analysis of the changes in Test 2 scores as compared with the Test 1 scores.

Turning first to the data for the younger *Ss* (groups 1, 2 and 3) the following observations may be made: (1) the mean increases are largest for the youngest *Ss* (4.7 points) and grow progressively smaller with increasing maturity (Table 2); (2) the range of score changes has to be extended to 36 points for the youngest group, to 30 points for group 2 and to 24 points for group 3 (Table 3);[12] (3) the majority of *Ss* in these younger groups gained or lost more than 3 points in their retest (Table 3).

TABLE 3
Percentage of cases within each ±3-unit span of variation of T2 scores from T1 scores (from Stanton and Koerth)

Span of Variation	Group 1	Group 2	Group 3	Group 4
±0	6.4	5.3	6.6	5.7
±1–3	24.6	44.2	38.8	44.6
±4–6	26.0	25.0	23.0	28.7
±7–9	16.5	12.5	19.1	12.7
±10–12	11.6	6.2	8.5	8.3
±13–15	6.0	4.8	1.3	
±16–18	2.8	0.5	0.7	
±19–24			2.0	
±19–30		1.5		
±19–36	6.1			

Stanton and Koerth have converted the mean raw scores into centile ranks. The mean score for the pre-adolescents in Test 1 was 76.5. Based on Seashore's standards for the fifth grade, this would be equivalent to a mean centile rank of 73.5. By the time these *Ss* took Test 2, however, their raw scores had to be converted on the basis of eighth grade standards, so that the mean raw score of 81.2 was equivalent to a rank of 69.9—3.6 centile units *lower* than the Test 1 rank.

[12] Graphs presented by the authors indicate that "there is a noticeable tendency for larger percentages of children to show a gain rather than a loss" (34, p. 12).

Similarly, the mean centile rank for the adolescents in Test 1 was 71.1 when interpreted on the basis of eighth grade standards, but the Test 2 mean for the same individuals three years later, interpreted according to the adult norms, was equivalent to a rank of 66.7—4.4 centile units lower. The results in both tests for the Group 3 *Ss* were transmuted according to adult standards and showed a gain of 2.8 points.

The increases in mean raw score and the general shift of the scores in a markedly positive direction for Test 2 are not attributed by Stanton and Koerth to musical training, but to the progressive lessening of cognitive factors with maturation (34, p. 19):

> A child's inability to express his finest sensitiveness of response may be due to several causes such as clumsiness in observation, limited span of attention, difficulty in following a task, excitement due to novelty, lack of controlled attention, et cetera. Educational development improves these factors within the range of each child's potentialities; consequently the difficulties in measurement are less as a child matures. In other words, we can only approach a child's 'proximate physiological threshold of hearing' in these measurements as the child matures. As the cognitive threshold more nearly approaches the approximate physiological threshold known only at maturity, the scores in the tests increase. At maturity, when the musical capacity scores tend to remain constant, the approximate physiological threshold is revealed. From the data at hand the writers are suggesting the eleventh grade in day school, or 16 years, as the age of maturity when development for the average child has been sufficient to enable him to reveal the approximate physiological threshold in hearing as measured by such musical capacity tests as were used in this study.

Since the gains for younger *Ss* are attributed to reduction of cognitive factors rather than to the musical training received, and since it is concluded that adults can reveal their approximate physiological limit in the first test, the results for the music degree *Ss* are of particular significance in this study.[13] Tables 2 and 3 indicate that (1) the mean increase for this group was only 0.1 point after three years of intensive musical instruction; (2) for this group, the Test 2 scores varied from the Test 1 scores a maximum of 12 points plus or minus and the authors state that the amount of loss almost balanced the amount of gain; (3) about half of the *Ss* in this group fluctuated no more than 3 points from their Test 1 score.[14] A variation of 3 points is regarded by Stanton and Koerth as a normal variation "within the natural fluctuation of attention and to be expected in measurements involving the hearing threshold" (33, p. 6).

The constancy of their mean score and the fact that so large a percentage of these *Ss* fluctuated no more than 3 points from their Test 1 score are interpreted as evidence that when they are competently given to intelligent adults, the Seashore tests measure physiological limits of capacities:

> The ultimate proof of the capacity nature of any test probably can never be found but these experiments show that the Measures approach the ideal of being measurements of musical capacities (35, p. 28).

It is possible to obtain from adults a meas-

[13] Ruth C. Larson, (9) examined the test results of music majors at the University of Iowa School of Music. The instruction received was similar to that for Stanton and Koerth's music degree *Ss*. There were three small groups of *Ss* who respectively had one, two and three years of musical instruction between Tests 1 and 2. Their mean gains in the pitch test were 0.03, 1.60 and 1.67 points. Larson concludes that inasmuch as these small mean gains are all within the P.E. of the score, instruction in theoretical and applied music did not lead to any significant gain.

[14] The test-retest reliability coefficient for the music-degree *Ss* was .54 ± .04. For Groups 1, 2 and 3 respectively, the reliabilities were .54 ± .03, .40 ± .04 and .64 ± .03.

ure of their greatest degree of capacity which will not vary significantly with musical training and education. . . . Shall we then say that musical capacities as measured by these tests reach a certain degree when one becomes an adult with no significant variation after that time? Evidence herein presented substantiates an affirmative answer . . . (*ibid,* p. 32).

Interpret as the reader may, the fact remains that if these Measures did not come somewhere near measuring native capacities there should be much greater gains observable in test scores of students after three years of intensive musical training and education (*ibid,* p. 39).

Discussion. The data presented by Stanton and Koerth relative to the effects of musical instruction on Seashore test scores are based *only on the test performance of students who remained in the Eastman School for at least three years,* for some of the students who were present for the first test had dropped out or were dismissed before the expiration of the three year period. The *Ss* whose scores were reported in these investigations did not fairly represent the population generally, therefore, but only a relatively homogeneous sampling of *Ss,* since (1) they had to have sufficient musical talent to be acceptable as students in the Eastman School of Music and (2) they had to survive academically in the School for three successive years. In the case of the music degree adults, these factors of selection were most pronounced.

This matter of selection raises a question as to whether the sampling of adults was sufficiently typical of adults generally to warrant the following statements of Stanton and Koerth (34, p. 6):

This constancy of scores provides objective evidence for two important facts: first, it is possible to measure an adult's greatest degree of musical capacities when the tests are given to adults for the first time under controlled conditions by an experienced examiner; second, the scores in the tests do not vary significantly after musical training and education. The first fact really means that an adult with average intelligence and natural educational development has the ability to give evidence of his psycho-physical limit of hearing and his sensitiveness of response to such stimuli.

With respect to this problem of sampling, it seems proper to consider the further possibility that if retest scores had been available for those students who dropped out or were dismissed before the retest was given, the mean score, even for the music degree adults, might have shown a greater increase. Support for this possibility is contributed by two lines of evidence: (1) Stanton has shown in another study (35) that students with low talent profiles were very short-lived in the School[15] and it may be assumed that at least some, if not most, of the low talent profiles included low ratings in pitch discrimination; (2) there is a tendency for initially high *Ss* to score lower in a retest, while initially low *Ss* showed more marked increases in the retest score.[16]

In arriving at the conclusion that the increase in scores of younger *Ss* was due

[15] None of the students whose talent profiles classified them as "discouraged" remained in the School past the freshman year and only 19.1% of the "doubtful" students entered the senior year. Moreover, among the students of one of the entering classes, the percentages of dismissals (mostly for academic reasons) were: 3.5% of the "safe"; 15.0% of the "probable"; 17.5% of the "possible"; 52.4% of the "doubtful" and 63.3% of the "discouraged" groups.

[16] Analysis of Wright's data, presented on p. 9, showed this trend even when no training was given. The same pattern is consistently found in data on the three younger groups studied by Stanton and Koerth. In groups 1, 2 and 3 respectively, Q1 was 0.4, 1.1 and 2.5 points *lower* in Test 2 than in Test 1 despite the fact that *Ss* were three years older and had had three years of intervening musical instruction. In these same groups, however, Q3 was 8.0, 5.1 and 5.6 points *higher* in Test 2 than in Test 1. It may be assumed that a similar trend would have been found for the music degree *Ss.*

to lessening of the cognitive difficulties rather than to musical instruction, Stanton and Koerth have relied principally upon the fact that the mean of the music degree adults was so stable, notwithstanding the more intensive musical instruction which they received (34, pp. 18-19):

> The question immediately arose, was this increase in scores due to musical training? The amount and quality of training were similar for each of the pre-adolescent, adolescent, and post-adolescent groups, and much more extensive and intensive for the adult group; yet the mean increases in scores were less as development advanced. It would seem, then, that amount and quality of musical training had little or no effect on scores in the Seashore tests.

If the interpretation of the stability of the mean for the music degree adults is tempered by recognition of the fact that these *S*s represented a very highly refined selection within a narrowly selected group, these adults become a less adequate criterion by which to evaluate the score increases of younger *S*s who were less highly selected. Thus, the conclusion of Stanton and Koerth that the score increases of younger *S*s were entirely due to "mental maturation accompanied by greater finesse in the functioning of cognitive factors" (*ibid.*, p. 18) is not satisfactorily established by these studies.

Although in their retests, relatively large percentages of *S*s fluctuated no more than 3 points from their Test 1 scores (31.0%, 49.5%, 45.4% and 50.3% of Groups 1-4 respectively), analysis of the authors' graphs shows that the Test 2 scores were substantially higher (7 or more points) in the case of at least 35% of the Group 1 *S*s, about 20% of the *S*s in Groups 2 and 3 and about 10% of the Group 4 *S*s. Gains as large as 7 points in raw score may represent rather significant changes in centile rank, particularly if the Test 1 score is equivalent to a rank in the neighborhood of the median, e.g., for an adult, a raw score of 81 is equivalent to a rank of 50, but an increase of 7 raw score points elevates the rank to 90.

Finally, it should be noted that even if all of the *S*s had fluctuated little or not at all in their retest scores, these studies could indicate merely that pitch discrimination, as measured by the 1919 form of the Seashore pitch test, was not susceptible of improvement through general musical instruction. They should not be used to imply that *other* types of training would result in the same absence of improvement. Any interpretations or conclusions drawn from these investigations should take cognizance of the fact that they reflect the effects of but one kind of "training," viz., musical instruction, and that this is not necessarily the *best* type for a crucial test of the "capacity" hypothesis. The present writer has elsewhere (45) suggested that such instruction may, in fact, be only remotely related to the "specifics" measured by the Seashore tests.

This possible irrelation of general musical instruction to "specifics" occurred to the writer as an extension of a point made by Seashore in a recent article (25) dealing with proper and improper criteria for validating the Measures. Seashore has protested against the use of validity criteria such as grades in applied music or music theory courses on the ground that although the *tests* measure "specifics," these *criteria* are "omnibus." His tests, he says, "represent the theory of specific measurements insofar as they conform to the two universal scientific sanctions on the basis of which they were designed; namely, that (a) the factor under consideration must be isolated in or-

der that we may know exactly what it is that we are measuring; (b) the conclusion must be limited to the factors under control" (p. 25). The only fair validity criteria, according to Seashore, must therefore also be "specific," e.g., for pitch discrimination, correlation with "objective records of musical performance in *pitch intonation* or ability to hear *artistic pitch deviation* in the musical situation" (p. 26). In most of the attempts to study the validity of the Measures, however, investigators have correlated scores in each of the tests against grades in music courses or against ratings of musical talent. Seashore maintains that this type of validation has little or no significance:

> They [the Seashore Measures] should not be validated in terms of their showing on an omnibus theory or blanket rating against all musical behavior, including such diverse and largely unrelated situations as composition, directing, voice, piano, violin, saxophone, theory, administration, or drums; because there are hundreds of other factors which help to determine job analysis in each of such fields.
> . . . I have been bombarded all these years by the omnibusists for this type of validation, but have persistently refused [action] on the ground that it had little or no significance (pp. 25-26).

It is clear, however, that the musical instruction received by *Ss* in the studies reviewed in this section was *also* "omnibus" in character. Such instruction may, therefore, have been just as "diverse and largely unrelated" to the tests as the criteria which were rejected by Seashore for validation! If the only acceptable criteria for validation of the tests are "specific," then "specific" training, closely related to the content of the tests, may also be necessary for proper evaluation of the "capacity" hypothesis. If we assume that the tests are valid, and if we accept Seashore's repudiation of "omnibus" criteria to challenge the validity of his tests, it seems proper to question also the use of "omnibus" training to demonstrate unimprovability.

It has been assumed in the studies here reviewed that when similar test scores are made before and after general musical instruction, it is conclusively established that a test measures a fixed, native "capacity." This assumption appears subject to challenge on the ground that such instruction may not be sufficiently related to the content of the individual tests to serve as "final critical proof" that the tests measure "physiological limits" of "capacities." It is doubtful whether anything short of training designed to be intensive and remedial and directed toward improvement of the specific "capacity" will fulfill the necessary requirements.

EFFECTS OF REMEDIAL TRAINING

The seven studies reviewed under this heading cover a period from 1903 to the present. Aside from the fact that all of the investigators gave individual training, there is little similarity in the methods used in these studies. The following training procedures are most frequently indicated: (1) verbal assistance including definition of pitch and its differentiation from other aspects of tone; suggestions that *Ss* employ analogous imagery, as for example, a visual image of a ladder; emphasis upon attentive listening; (2) demonstrations with foreknowledge of the correct answers; (3) informing *Ss* as to the correctness of their responses; (4) drill in interval recognition; (5) practice in vocal matching of a single tone, intervals or tonal sequences.

The majority of these studies were conducted with adults as *Ss*, but data on

children are to be found in two of the seven studies reported here. Although there seems to be little point in employing highly proficient *Ss*, only four of the seven investigators made an attempt to select *Ss* with deficiency in pitch discrimination or in singing ability.

In only two of these seven investigations were scores or ranks in a standardized test employed as criteria for measuring the effects of training. Tests given by means of tuning forks were most frequently employed to determine improvability, but an oscillator, a Stern tone-variator and even a piano were also used for this purpose.

G. M. Whipple. In 1903 Whipple (38) reported on the pitch discrimination of one adult who gave considerable evidence of being unmusical. In childhood this *S* had been told that she would never be able to sing, at the age of fourteen 'didn't even know how to try' and was excused from participation in a school chorus. As a college senior she was still unable to sing and even her whistling proved to be so inaccurate that it would have been difficult to recognize the tune were it not for the rhythm. Her perception also indicated deficiency, for she could not distinguish any difference between a major and a minor triad or detect a change of a semitone in a familiar melody. When presented with semitone differences played on the piano at three different frequency levels, only 40%, 74% and 70% right answers were given. With the Stern tone-variator this *S* "was frequently unable to judge correctly" a difference of 12 ~ at a standard of approximately 250 ~.

"Systematic drill and coaching" (not described as to nature or extent) "rapidly increased the discriminative sensitivity," for the threshold (78% right answers) with the tone-variator stimuli was reduced to 2.8 ~. This improvement did not transfer very effectively to discrimination of piano tones, however, for only 68%, 68% and 78% right judgments were given at the three frequency levels respectively.

Whipple remained in doubt as to the effectiveness of the training. The lack of transfer from variator tones to piano tones tended, he thought, to strengthen the idea that an individual may be "constitutionally and inevitably unmusical." On the other hand, the "rapid daily rise of efficiency" with the variator tests led Whipple to believe that a longer period of training might have brought about more definite and permanent improvement.

The lack of transfer from improvement in discriminating tone-variator stimuli to discrimination of piano semitones suggests the need for further investigations along these lines. It is important to know not only whether individuals can improve their performance in a specific pitch test, but also whether improvement, if it occurs, will spread to tests which employ different frequencies, different timbres and different psychophysical methods.

F. O. Smith. Using tuning forks and resonators as recommended by Seashore (17), Smith (31) gave two tests of pitch discrimination to a class of 200 adults. The poorest one-fourth of the group, 54 in number, then received "personal instruction" described as follows: "An effort was made to find out what particular difficulties they were encountering, and explanation and illustration were based progressively upon this information" (p. 73).

All but seven cases made "rapid improvement." The thresholds for the 47

Ss who improved are shown in the accompanying distribution:

Thresholds	*30 ~*	*23 ~*	*17 ~*	*12 ~*	*8 ~*	*5 ~*	*3 ~*	*2 ~*	*1 ~*	*0.5 ~*
N before instruction	6	2	8	8	12	11				
N after instruction		1	1	1	5	11	12	5	7	4

It is noteworthy that before instruction, not a single *S* in this group had a threshold under 5 ~, but that after receiving instruction, 60% had thresholds of 3 ~ or less. It is also interesting to observe that the number of *S*s with high thresholds (12 ~ or greater) was reduced from 51% before instruction to only 6% after instruction. Smith concluded that all tests should be preceded by efficient instruction, preferably individual, and that "all who show poor records must be subjected to more intensive and searching instruction before the record can be accepted for serious purposes" (p. 75).

In another portion of Smith's investigation, 106 elementary school children with high thresholds in the group tests were given "individual practice" in order to aid them to "distinguish different tone qualities and to form right habits of attention" (*ibid*). This "special assistance" was given concurrently with a series of twelve group tests. For the 71 boys, the average threshold in these tests was reduced from 17.3 ~ to 9.8 ~, while that for the 35 girls dropped from 17.7 ~ to 7.8 ~. Although these final thresholds are still quite high, the fact remains that with what must have been relatively brief and superficial training, these children greatly improved their preliminary records.

One of Smith's most valuable contributions in this research is his qualitative analysis of developmental factors in pitch discrimination. These consist essentially of three types of "habits of control:" (1) auditory and kinaesthetic sensations; (2) auditory, visual and kinaesthetic memory images; (3) special attitudes, such as feeling of familiarity, most favorable form of attention, interest, etc. Following is a digest of some of these developmental factors, gleaned from introspections of *S*s:

1. Some *S*s perceived pitch chiefly in terms of tonal qualities, but the particular quality varied with different individuals. The lower tone was often distinguished from the higher as being duller, deeper, heavier, more mellow.

2. A large number of *S*s depended upon kinaesthetic sensations in the vocal organs. Some said that they could not tell whether the second tone was higher or lower until they reproduced the tone, either audibly or mentally. One adult who had failed to distinguish a difference below 20 ~ tried humming the tones. He immediately reduced his threshold to 8 ~ and subsequently to 2 ~.

3. A variety of other kinaesthetic sensations were reported, e.g., a tendency to move up and down with the tones, to breathe more deeply for the lower tone, etc., but in most instances, auditory and kinaesthetic sensations were combined into a single experience.

4. Many *S*s reported carrying over an auditory image of the first tone for comparison with the second. Visual imagery included localization in space with the higher tone usually imaged above the lower. Some referred the tone to a particular musical instrument and thought of how they would play the higher and lower tones.

5. Some *S*s listened to the beginning of

the tones, others to the end and about half of the *Ss* reported that they listened to the middle portion of the tones. Being set for a definite portion of the tone seemed to lead to more rapid judgments and, when brought under control, this seemed to favor improvement.

6. It was reported by some *Ss* that the closest attention was required for successful discrimination. Others, however, found that this caused nervous strain which led to mistakes. Smith found this to be true in his own case and one of his *Ss* (TFV) reported: "Much depends upon my attitude. If I hold myself in a passive attitude and answer with ease, in a reflex way, I am quite sure to be correct in my judgment; but if I get the attitude of strict attention, I cannot do so well. If I can keep in a state of relaxation, I experience no difficulty in giving the judgments" (p. 91).

7. As skill in pitch discrimination developed, all of the above factors usually tended to become mechanized and *Ss* began to grasp the interval as a whole with no awareness of the factors which entered into the judgment.

Although Smith, following Seashore, characterized the factors which impeded maximal performance as "cognitive," they seem more consistent with R. H. Seashore's work methods hypothesis. Apparently the factors which hindered optimal discrimination were not merely distraction, failure to understand instructions, lack of good will, etc., for analysis of verbal reports of *Ss* showed that there were other factors of importance, viz., learning to reproduce the tones vocally or subvocally; utilizing auditory or kinaesthetic associations; sharpening of imagery; learning the optimal adjustment of attention, etc. The data also indicate that whatever the factors may have been which prevented the immediate disclosure of the "physiological limit," they were not overcome by simple retesting, for many of Smith's *Ss* continued to improve even up to the twelfth test.[17]

E. H. Cameron. This experiment (1), reported in 1917, was concerned in part with the effects of practice in singing a tone of a certain pitch on ability to discriminate (1) tones at the same pitch level (2) tones at a different pitch level. The six psychologists who served as *Ss* all received pre-training and post-training tests in pitch discrimination at two frequency levels, 100 ~ and 225 ~. In addition, *Ss* were given pre-training and post-training tests of their accuracy in vocal reproduction of these same two frequencies.

For the tests of pitch discrimination, the following method was used: Four tuning forks were employed—two standard forks with frequencies of 100 ~ and 225 ~ respectively and two comparison forks fitted with adjustable weights.[18] The forks were electrically excited, the sounds reaching *Ss* in another room through a telephone arrangement. Judgments were classified as (1) higher; (2) lower; (3) same or doubtful. Two periods of about 20 minutes each were allowed for adaptation before any responses were recorded. For each *S* the daily threshold was the smallest difference for which five successive correct judgments were given and the final threshold was the average of five such daily thresholds for different days.

No direct coaching in pitch discrimi-

---

[17] *Vide* p. 8, this paper.

[18] Qualitative differences are almost inevitable in forks of this kind (17) and these admittedly constituted a source of error in this experiment.

nation was given, but several months of daily practice in vocal matching of a tonal stimulus intervened between the initial and final tests. Three of the Ss—M, R and My—were given practice in singing the 100 ~ tone, while the other three—A, F and Ay—were practiced on the 225 ~ tone. Twenty trials were made daily, making a total of approximately 1000 attempts for each *S*. Graphic records were made of these sung tones. After this rather intensive practice, it was found that four of the Ss, F, Ay, M and R., succeeded in greatly reducing their average error in vocal matching of the practiced tones. For these Ss there was little evidence of transfer to accuracy in singing the unpracticed tone, however. Three of them reduced their error slightly but one made a poorer record. Of the other two remaining Ss, one did not succeed at all in approximating the standard and the other was erratic, matching the standard only when he sang with the tuning fork sounding.

At the conclusion of this vocal practice, Ss were retested in pitch discrimination. It was found that the four Ss who had reduced their errors in vocal reproduction of a tone also achieved lower thresholds in discrimination of tones *at the same frequency level* at which their singing practice had been given:

S	*pre-training threshold*	*post-training threshold*
F	2.2 ~	1.1 ~
Ay	2.2 ~	1.3 ~
M	1.8 ~	0.6 ~
R	2.6 ~	1.4 ~

At the *unpracticed* level, however, the thresholds of discrimination for these same Ss changed very little:

S	*pre-training threshold*	*post-training threshold*
F	2.6 ~	2.4 ~
Ay	2.1 ~	2.0 ~
M	1.4 ~	1.8 ~
R	1.3 ~	1.2 ~

In the case of the other two Ss who had made no real progress in matching tones, post-training thresholds remained about the same as pre-training thresholds at both the 100 ~ and the 225 ~ standards.

To summarize: Those Ss who improved in accuracy of singing also improved in pitch discrimination at the same standard frequency, while those who did not improve in singing accuracy did not improve in pitch discrimination either. But even the Ss who improved in accuracy of both intonation and discrimination at one frequency level did not significantly transfer this improvement to intonation or discrimination of tones at the other frequency level.

Cameron's investigation sheds some light on the problem of the cognitive limit. The six Ss in this study were all trained psychologists for whom such cognitive obstacles as lack of application, failure to understand instructions, distraction, poor motivation, etc., would surely be at a minimum. Moreover, Cameron allowed two 20-minute periods of practice before taking records. It may be assumed further that if cognitive obstacles were present, they would have operated to just as great a degree at the level at which no singing practice was given as at the practiced level. Cameron's results show, however, that when there was substantial improvement, it occurred only at the practiced level and did not transfer to the unpracticed level.

Taking issue with Seashore, Cameron ascribed improvement in discrimination to the practice in singing rather than to adaptation, interest, attention or other similar cognitive factors. He suggested that even such a relatively simple process as sensory discrimination of tones depended upon the "organic unity" of motor and sensory factors and concluded that . . . "With the development of more

precise and unvarying modes of response to one tone, there arises a greater keenness in discriminating that tone from all others" (p. 179).

Cameron's results imply that for improvement to take place throughout the tonal range, it may be necessary to practice motor reproduction of tonal stimuli over a similarly wide range. Since individuals are so limited vocally, such practice might be given by permitting *Ss* to manipulate some instrument with variable pitch. Moreover, with the aid of modern apparatus, such as the Conn Chromatic Stroboscope (47), *Ss* could readily obtain a visual check on the accuracy of their own intonation. It is possible that this would prove a more effective training procedure, since *Ss* could tell at a glance both the direction and the extent of their deviation from the standard tone.

M. Wolner and *M. Wolner and W. H. Pyle*. Wolner's experiment, conducted in 1932 (39, 40), is a study of the effects of intensive and diversified training upon the singing ability and the pitch discrimination of seven children with extraordinary initial deficiency in singing and in perceiving pitch differences. These *Ss* were selected by eliminating all but the seven most extreme cases from among a larger group of pitch-deficient children. None of these seven *Ss* could sing although they had all been in music classes since the first grade and were in grades 5, 6 and 7 at the time of the experiment; they could not discriminate piano tones even when the differences were as large as a fifth or an octave; they were unable to distinguish tuning fork differences as large as 30 ~ at a standard of 435 ~.

The training took several forms. Verbal assistance was given by *E* in his attempt to define pitch, differentiating it from loudness, duration and timbre; in his use of the analogy of a ladder and a musical scale as an aid to visual imagery; in his suggestion to *Ss* that they think of tones as one would think of a problem. Remedial training with the piano consisted of (1) vocal reproduction of tones played on the piano, at first single tones, but later short sequences and, as skill developed, diatonic and chromatic scales and intervals; (2) drill in the discrimination of intervals, reverting to singing methods when wrong answers were given. Wolner also employed tuning forks for portions of his training. A standard fork of 435 ~ was used with comparison forks which were higher than the standard by 30, 23, 17, 12, 8, 5, 3, 2, 1 and 0.5 ~. When the three largest increments were played, the children were asked to reproduce the tones vocally, sometimes singing the words "low" and "high" for the lower and higher tuning fork tones. The training was given individually, each child receiving 20 minutes of training and testing each morning, five days a week for 81 days—an average of sixteen hours. Great patience and perseverance were reported necessary, particularly in the early weeks of training.

In the tests with tuning forks, Wolner struck each fork with a mallet and held it to a resonator for two seconds, allowing an interval of one second between the two tones of a pair. The "standard of perfection" used was two sets of ten successive correct responses. This was termed "achieving" an increment.[19]

After training, all seven of the *Ss* improved remarkably in singing ability[20]

[19] According to the Wolner thesis (39), the criterion for "passing" the tuning fork tests was 100% correct answers in two sets of ten trials. According to the article by Wolner and Pyle (40), however, this criterion was 100% correct answers in only ten trials.

[20] One child succeeded in singing, without pitch deficiency, major and minor scales, chro-

and succeeded in discriminating without error all piano intervals, including semitones, over a four-octave range. Four of the *Ss* "achieved" all tuning fork increments from the largest difference of 30 ~ down to the smallest of 0.5 ~. With the same criterion for "passing" a test, viz., two sets of ten successive correct judgments, one of the three remaining *Ss* succeeded with pitch differences from 30 ~ to 2 ~, while the other two "achieved" increments of 3 ~ and 8 ~ respectively.[21]

For most of the *Ss*, the number of tests and the amount of training time necessary for "achieving" an increment seemed to conform to patterns often found in experiments on learning of skills. The greatest difficulty was frequently encountered with the 30 ~ and the 23 ~ increments. Once these were mastered, however, a "spurt" of improvement followed which extended for several more difficult increments, only to be followed by another period of seemingly arrested progress at an increment which required another delay before it could be mastered. Thereafter, however, more and more difficult increments were apt to be "achieved." An account of J.P.'s progress illustrates this tendency (39, p. 19):

. . . it took him four weeks to conquer the 30 dv. fork. Following this, he passed the 23 dv., 17 dv., 12 dv., and 8 v., forks with comparative ease. Upon reaching the 5 dv. fork, he experienced slight difficulty. After a week of training, he passed it and then went to the 3 dv. fork, on which he remained two weeks. From this he proceeded to the 2 dv. fork, and remained at it for a whole week. Here he had a little trouble. In both the 3 dv. and 2 dv. forks, the influence of intensity was markedly apparent. Upon completing the 2 dv. fork test successfully, he next advanced to the 1 dv. and .5 dv. forks respectively, which he achieved without effort.

It was concluded that (1) the pitch deficiency of these seven children was due to some failure in method rather than to any anatomical defect of the inner ear or any neural derangement; (2) defective pitch discrimination was remediable through systematic and adjusted instruction and practice; (3) most pitch-deficient children can probably be trained to distinguish pitch with considerable accuracy.

The advantages of individual training over group training are clearly brought out in this study. In the handling of groups, the training may be given at levels which are too easy for some and too difficult for others. In the individual method, however, the chances for effectiveness are greater, since *Ss* can be trained at any one time on that increment at which the training is deemed most efficient. As Wolner indicates, there is no point in commencing training with a 30 ~ increment when *Ss* cannot discriminate a difference of an octave. Individual differences in response to method or changes in method constitute another factor which is best met by individual training. Wolner has suggested that constant "changes, innovations, variations, and shifts" were necessary for the success of the experiment (39, p. 44). Moreover, in working with an individual, *E* can adapt to his particular requirement the amount of time needed by *S* to reach a certain level of proficiency

---

matics, intervals, tones picked at random and several songs with words; another learned to sing scales, intervals and the tune of a song without words; two children were able to sing scales and intervals; three sang scales with great improvement, but not perfectly.

[21] The *S* who "achieved" the 8 ~ increment according to the Wolner thesis (39) is reported in the Wolner-Pyle article (40) to have achieved" a 5 ~ increment.

and can add more difficult increments to the training procedures when and only when the larger increments are mastered. In short, if training is to be properly remedial, it should take cognizance of the unique problems presented by each individual.

In considering the educational implications of this study, it should be noted that the time factor may be very important. A pitch-deficient individual may not be able to discriminate a difference of 30 ~ after a few hours or even a few days of training, but may require months before the training becomes effective. Even after 81 days of training, Wolner was not certain that he had reached the "physiological limit." On the contrary, it was believed that the three poorest *S*s might have improved still further if the experiment had been prolonged.

Consistent with the Cameron investigation, it appears probable that in effective remedial training, *S* is given ample opportunity for *active* experience in dealing with the stimuli. Mere verbal explanations and even a long period of listening may not be adequate to effect improvement, while extended training in vocal reproduction of tones may be highly effective. Wolner and Pyle believe that this singing experience broadens the pupil's conceptions of high and low pitch to the extent that he feels the muscles of his vocal organs tightening or relaxing. This process, they believe, allows the pupil to demonstrate to himself in a practical way the meaning of pitch differences.

The following modifications and extensions of this study might be suggested: (1) more precise control of the test stimuli by the use of other apparatus or a standardized test (manually excited tuning forks are known to present numerous sources of error); (2) employment of a more reliable criterion of achievement; (3) determination of the permanence of the improvement; (4) determination of the degree to which improvement might transfer to other frequency levels and other types of tonal stimuli; (5) a similar experiment conducted on adult *S*s.

R. H. Seashore. The *S*s in this experiment (29, 30) were twelve adults who were selected because of their poor showing on two consecutive administrations of the 1919 form of the recorded Seashore pitch test. The averages of these two test scores ranged from 58 to 68, equivalent to centile ranks of 12 or less. After individual training, nine of these *S*s were given two retests. Initial and final pitch discrimination tests were also given with a beat-frequency oscillator.

The oscillator, used in connection with a dynamic speaker, was employed in the training, which was given in weekly periods of 45 minutes. The time devoted to the training ranged from 3 to 9 hours, with an average of 5.6 hours. The standard frequency was the same as that used in the Seashore test (435 ~) and the general technique was also similar. Training procedure included (1) demonstrations on easily noticeable differences with knowledge of what was to come each time and (2) informing *S* each time he made a judgment as to whether it was right or wrong. After the first period, most of the time was spent in practice slightly below the most recently determined threshold, i.e., at what was thought to be the most efficient level for each *S* at any given time.

The results may be considered in terms of (1) initial and final thresholds in the oscillator tests (75% correct judgments on at least 50 trials) and (2) initial and

final centile ranks in the Seashore pitch test. In the oscillator tests, the mean threshold dropped from 9.2 ~ to 4.6 ~. Ten of the twelve *Ss* improved in these tests and seven of them achieved post-training thresholds of 3 ~ or less.[22] All of the nine *S*s who took the final Seashore tests improved. While the average initial rank for two testings was only 6.6, the average rank in the two final testings was 45, with three of the *S*s well above the median for the adult population.[23]

On the basis of these results, it appears that in many instances the pitch discrimination of seemingly pitch-deficient adults is improvable with relatively little training. A longer period and a greater variety of training methods might have resulted in still further improvement. These results are significant, however, in pointing out the possibilities.

From this study it appears that training given with one type of source (oscillator) transferred effectively to another type (recorded tuning fork tones) at the same standard frequency. A few incidental experiments with the oscillator seemed to indicate that there might be some transfer to adjacent tonal regions, but no formal quantitative proof along these lines was reported.

R. H. Seashore has suggested that another experiment be performed in which suggestions for improving work methods be used as training devices: "The crucial experiment will be to determine whether individuals who are instructed in superior work methods can significantly improve their own thresholds" (30, p. 129).

A. A. Capurso. The major problem in Capurso's experiment (2, 3) was to determine whether an "associative technique" might prove effective as a means of improving pitch and interval discrimination. In one portion of this experiment, seven experimental *S*s and six control *S*s participated. These individuals were selected from a group of 58 adult music students on the basis of their scores in one administration of the Seashore pitch test (1919 form). Of the thirteen individuals selected, the seven highest had scores ranging from 87 to 92, while the six lowest ranged from 50 to 84. Four high cases and three low cases were chosen for the experimental group, leaving three high and three low *S*s for a control group.

The experimental *S*s received an average of 10.5 hours of individual training distributed in 30-minute periods on alternate days of the week and extending over a period of seven weeks. This training consisted of drill in interval recognition and in pitch discrimination. Intervals of fifths, fourths, thirds, etc. were played on the piano and *S*s were asked to form an association with some other auditory stimulus or with some "mood word." While fifths were played, *S*s tried to associate the auditory effect of the interval with the ringing of chimes or church bells. The bugle-call "Taps" was associated with fourths. For certain other intervals, *S*s were encouraged to form an association between the sound of the interval and a "mood word" such as "tumult," "longing," "comfort," etc. The technical musical names for the intervals were later substituted for these first associations.

After this drill in interval recognition, Capurso's *S*s were trained in discrimination of tuning fork tones. No indication is given, however, of the amount of time

[22] Initial thresholds were as follows: 28, 28, 8, 8, 8, 8, 8, 8, 5, 5, 3 and 3 ~. Final thresholds for the same individuals respectively were 17, 0.5, 8, 5, 5, 3, 3, 3, 3, 1, 5 and 2 ~.

[23] Initial ranks were 12, 9, 8, 7, 6, 5, 4, 4 and 4. For the same individuals, final ranks respectively increased to 29, 48, 11, 81, 91, 70, 40, 19 and 17.

allotted to this type of training. Tuning forks of similar frequencies to those in the Seashore test were used, but Capurso altered the Seashore method in the following two respects: (1) the differences in pitch were achieved by using various combinations of tuning forks, e.g., a tone of 465 ~ might be followed by another of 435.5 ~ while in the next trial a tone of 436 ~ might be followed by a comparison tone of 458 ~ etc.; (2) instead of asking *Ss* to tell whether the second tone was higher or lower than the first, as in the Seashore test, Capurso asked for a judgment as to which of the two tones was higher. During the training, *Ss* were informed as to the correctness of their responses.

A retest with the Seashore record followed the seven weeks of training. Raw scores and centile ranks for the seven experimental *Ss* and the six control *Ss* are shown in Table 4 (modified from 2, p. 816).

TABLE 4
Scores and ranks of Capurso's experimental and control *Ss* before and after training

S	Scores Beofre	Scores After	Ranks Before	Ranks After
Exper.				
1	92	92	99	99
2	90	89	96	94
3	88	90	91	98
4	88	91	91	96
5	77	86	32	81
6	73	78	21	36
7	50	90	3	96
Mean	79.7	88.0	61.9	85.7
Control				
1	88	88	91	91
2	87	88	87	91
3	87	82	87	56
4	84	84	70	70
5	77	81	32	50
6	74	73	23	21
Mean	82.8	82.7	65.0	63.2

While five of the seven experimental *Ss* improved, only two of the control *Ss* showed gains. The experimental group as a whole gained an average of 8.3 raw score points and about 24 centile points, while the control group made about the same average in the retest as in the initial test.

It should be noted, however, that omission of the rather spectacular results for the experimental *S* whose score increased from 50 to 90, lowers the average gain for the other six experimental *Ss* to only 3.0 raw score points and changes the average centile rank in the post-training test only about 12, rather than 24, points. Other factors which should be considered in interpreting this portion of Capurso's experiment are the following: (1) there are really only three *Ss* whose results are significant in the present problem, for the control *Ss* received no training and four of the experimental *Ss* were either in or very close to the highest decile of the adult population even before the training was begun; (2) the failure to give a pre-training retest, particularly to *Ss* who had low scores in the initial test, tends to diminish the significance of the results, for the reliability of low scores in a first testing is known to be doubtful; (3) from an empirical point of view, training in the recognition of intervals of fifths, fourths, thirds, or even of minor seconds, seems somewhat remote from the very fine discriminations required in at least half of the trials in the Seashore test.

The second portion of Capurso's experiment deals with two *Ss* who received training over a six-month period. Each of these *Ss* was given five testings with the Seashore record. The second test followed the forming of associations for fifths, fourths and sixths; the third test was given after *Ss* had formed their associations for the remainder of the inter-

vals; the fourth test followed that portion of the training in which technical musical names were substituted for the first associations; the fifth test followed an interval of training with tuning forks.

For one of the *Ss*, progressive scores in the five Seashore testings were 53, 68, 61, 68 and 71. Although this increase in raw score appears significant, it should be noted that equivalent centile ranks remained sub-average at 3, 12, 5, 12 and 17. Capurso suggested that diversion of interests and low motivation may have been responsible for the absence of more significant improvement.

The second *S* in this portion of the experiment showed more marked development. Scores in the five tests were 62, 80, 88, 87 and 89 with equivalent centile ranks of 6, 45, 91, 87 and 94—a change by stages from the lowest decile to the highest. Capurso reported that prior to training, this *S* had never been able to match tones vocally, but that after the third test she succeeded in doing so without difficulty and even learned to sing an ascending and descending scale without the aid of the piano.[24]

E. Connette. During a five-day period, Connette (4) gave individual practice in pitch discrimination to 23 adults, informing them as to the correctness of their responses. A standard tuning fork of 440 ~ was used with comparison forks which were higher than the standard by 30, 17, 5, 2, 1 and 0.5 ~ respectively. The forks were manually excited and each tone was allowed to sound for approximately 5 seconds. During each of the half-hour sittings, *Ss* heard each pair of stimuli 15 times, responding "higher" or "lower" after hearing the second tone. Immediately after the response, *S* was told whether it was correct or incorrect.

The group thresholds for the five days were: 7.33 ~, 6.51 ~, 5.25 ~, 5.20 ~ and 3.77 ~. The *Ss* who were in the upper half of the distribution on the first day made a 29% improvement, reducing their average threshold from 2.28 ~ to 1.62 ~. The *Ss* who were in the lower half of the distribution made a 58% improvement, reducing their average threshold from 12.69 ~ to 5.44 ~. In 20 of the 23 cases, the final thresholds were lower than the initial thresholds. Since the other three *Ss* had initial thresholds of less than 1 ~, their thresholds depended on so few trials that Connette regarded them as "doubtful cases."

Connette concludes from his data that it is evidently impossible to classify individuals permanently on the basis of a half-hour test. "Under favorable conditions," he says, "training effect is large and the individual in the poorest class originally may gradually move up to the best class in a week's time" (p. 529).

The validity of the distinction between the physiological and the cognitive threshold is questioned by Connette on the grounds that

> . . . The physiological threshold as a general limit of attainment is too broad a generalization from a single kind of experimental procedure. Even within one experimental situation there would seem to be no reason for assuming that learning approaches a limit in any kind of performance until that fact is

[24] Following is an excerpt from this *S*'s verbal report (2, p. 26): "When I practiced [violin] I would become discouraged and would give up in utter despair. Many people who play a little do not realize when they are wrong. And in my case, people would have to tell me when I was not playing correctly. But I was uncertain as to just what to do. If, for instance, I played a tone on the violin and tried to compare it on the piano, I could not tell whether they were the same or not. It seems I could not hear the tone at all; to me it was just a blank sound. . . . When I came to the University I was ashamed to say I had been taking lessons for so long a time and had accomplished so little. . . . Now, having had some training in pitch, I am beginning to hear the tones and am being [*sic*] able to discriminate especially well in the low register which was originally the harder for me to hear."

demonstrated adequately. It is so seldom possible to say with certainty that a limit has been measured in a series of experiments that . . . the concept seems of doubtful utility. If this is abandoned, there is no further use for the term 'cognitive threshold' either, as its main connotation is simply 'non-physiological.' (p. 531).

The "physiological limit" is regarded by Connette as "a function of a particular learning situation" and he holds that a new "limit" may result from a longer training series or another sort of procedure. He sees no cogent reason for regarding a non-improvable threshold as "limited by the characteristics of the receptor organ rather than by the conditions of any other parts of the system involved," and he is not convinced, he says, that the cognitive factors are the only ones which are subject to modification.

The following observations seem pertinent in evaluating the contribution of this research to the general problem: (1) when manually excited tuning forks are employed for pitch discrimination tests, the possibility that *S*s may discover cues, such as localization, timbre, intensity, etc., cannot be disregarded; (2) considerable improvement was shown by the group as a whole, but individual differences were not eliminated (the average threshold for the initially inferior *S*s remained poorer after training than that of the initially superior *S*s before training); (3) the time allotted to the training was extremely brief and the training technique was very limited; (4) Connette's conclusions with respect to the "physiological limit" are consistent with R. H. Seashore's "work methods" hypothesis.

Summary of data. Following is a brief résumé of the results presented in the various studies reported under the heading of remedial training:

1. After a short period of "drill and coaching," Whipple's one *S* improved greatly in the discrimination of tone-variator stimuli, but this improvement did not transfer to the discrimination of piano semitones.

2. In Smith's study of 54 adults, 47 (87%) showed marked improvement after "explanation and illustration." None of the 47 *S*s had low thresholds (3 ~ or less) at the outset, but 28 of them attained such thresholds after training. Similarly, the number of *S*s with high thresholds (12 ~ or more) dropped from 24 cases before training to only three cases after training.

3. With 106 children, Smith found that after instruction in "distinguishing different tone qualities and forming right habits of attention," the average threshold dropped from a value over 17 ~ to approximately 10 ~ for the boys and 7 ~ for the girls.

4. Of Cameron's six adult *S*s, four lowered their thresholds at the frequency level at which vocal practice was given from 2.2 ~ to 1.1 ~; from 2.2 ~ to 1.3 ~; from 1.8 ~ to 0.6 ~ and from 2.6 ~ to 1.4 ~ respectively. The other two *S*s did not improve. None improved at the unpracticed level.

5. All seven of the pitch-deficient children trained by Wolner improved. Before training, the children were not only unable to discriminate a 30 ~ difference with tuning forks, but they were even unable to discriminate piano tunes. After diversified and intensive individual training, four of the *Ss* "achieved" tuning fork differences of 0.5 ~ and the other three "achieved" differences of 2, 3 and 8 ~ respectively. All of them learned to discriminate piano tones successfully.

6. With twelve adults, R. H. Seashore found that after training with an oscil-

lator, thresholds in the tests given with the oscillator dropped from an average of 9.2 ~ to 4.6 ~. For nine of these *Ss* who completed the experiment, the average centile rank in the Seashore pitch test changed from 6.6 to 45. While all of the pre-training ranks were below 12, three of the nine *Ss* who participated in this portion of the experiment achieved post-training ranks of 70, 81 and 91 respectively.

7. Two of Capurso's initially high experimental *Ss* did not improve their Seashore pitch test scores, but the other five scored higher after training in interval recognition and pitch discrimination. The average score for all seven changed from 79.7 to 88.0 (equivalent ranks for these scores are approximately 62 and 86). The three low *Ss* respectively had raw score increases from 77 to 86, from 73 to 78 and from 50 to 90 with changes in centile ranks from 32 to 81, from 21 to 36 and from 3 to 96. The control group did not show improvement.

8. After six weeks of training, principally in interval recognition, one of Capurso's special *Ss* gradually increased her score in the Seashore pitch test from 62 to 89, equivalent to a change in centile rank from 6 to 94. The other special *S* changed his score from 53 to 71, but equivalent ranks of 3 and 17 make this change seem less significant.

9. After five half-hour periods of being informed as to the correctness of their responses, the average threshold for the 23 adults in Connette's study was reduced from 7.33 ~ to 3.77 ~. Three initially superior *Ss* did not lower their thresholds.

Evidence of improvement in pitch discrimination may be found in every investigation reported above. Although training did not eradicate individual differences and there were instances in which no improvement occurred, in general the data show reduction in the range of differences and a shift to a better level of performance.

IMPLICATIONS FOR RESEARCH

The following implications from the research literature seem pertinent in planning future investigations:

1. *Selection of Ss.* The pre-training status of *Ss* would be an important consideration in a crucial experiment. To lessen the possible presence of cognitive factors, it would seem preferable to select as *Ss* intelligent adults, who although they seem to possess a comprehension of the nature of pitch, make a poor record in a pitch discrimination test even when a retest is administered. If, in addition, *Ss* are attentive, highly motivated and practiced in test routine, the sample would seem even more desirable for an experiment along these lines. There would be no point in using *Ss* with high initial test scores, as there would be no opportunity for them to reveal improvement if it occurred.

2. *Number of Ss.* The larger the group and the more intensive the training, the more reliable the results are likely to be. If, however, there must be a choice between a small group which can be given intensive individual training and a large group which can be given only superficial training, the former would certainly seem preferable.

3. *A pre-training retest is important.* Although initially pitch-deficient *Ss* are desirable for an experiment of this type, such deficiency should in all cases be verified by a retest. Seashore has recommended this, and no investigation of improvability can be regarded as crucial if the retest is omitted. The most suitable

Ss are individuals whose scores remain low even after retesting, for the second test would presumably eliminate or greatly reduce the conventional cognitive factors. Moreover, in a practical situation, it would be this retest score which would form the basis for guidance.

4. *Sound source in the tests.* The guiding principles in the selection of the sound source are (a) accuracy and (b) constancy of all variables other than pitch. Ideally the sound stimuli would be pure tones of uniform intensity and duration, produced in some manner which eliminated all extraneous cues such as localization, noises from impact and damping, etc. Automatic production of the stimuli seems preferable. Although phonograph recordings for pitch discrimination tests are admittedly a makeshift, (26, pp. 307-308) the chief advantage in the use of a standardized test is that scores and ranks are used for guidance purposes and the data might therefore have important practical applications and implications.

5. *Time allotted to training.* With due caution as to the possibility of fatigue or monotony, the greater the aggregate number of hours and the longer the period over which the training is spread, the more crucial the experiment is likely to be.

6. *Training procedures.* The procedures should be sufficiently diversified to make them remedial for as many *Ss* as possible. It may be inferred from the literature that motor factors, manifested, for example, in vocal reproduction of tones, are important, at least for some *Ss* and at a certain stage in the training. The Tonoscope (18) or the Conn Chromatic Stroboscope (47) would seem to possess considerable potential promise in this respect, for with the use of such apparatus *Ss* could obtain a visual check of their accuracy. An ideal experiment would probably provide opportunity for active motor experience, preferably vocal or possibly manipulative, coupled with the aid afforded by the use of the stroboscope. In experiments along these lines, the training should be given individually and preferably under the supervision of an individual who is trained in both music and psychological experimentation.

7. *Quantitative factors.* (a) Data expressed solely in terms of central tendency often obscure highly significant changes, particularly if the group is comprised of many initially proficient *Ss*. (b) The basis for measuring improvement should be reliable and achievement should be based on an adequate number of trials. (c) Investigators have not used statistical techniques which would have subjected their data to critical tests of significance. The application of critical ratios or of the t test would seem advisable.

8. *The problem of transfer.* Seashore's implicit assumption of group factors in pitch discrimination has not been adequately explored. The transferability of the effects of training should be further investigated along with further research on the immediate problem of improvability.

9. *Qualitative aspects.* If pitch discrimination is improvable, it is important to inquire into the nature of the modifications which take place as *Ss* become more proficient. Careful qualitative analysis may lead to a clearer definition of superior work methods in pitch discrimination and to a better understanding of the limiting factors which block successful discrimination.

III. Objectives and Procedure in the Present Investigation

OBJECTIVES

The principal purposes of the present experiment were (1) to ascertain whether intensive training at one frequency level would improve pitch discrimination at the same level and (2) to determine whether, if there were improvement, it would transfer to pitch discrimination at other frequency levels.

PLAN OF THE EXPERIMENT

Throughout the experiment, an attempt was made to adapt the procedures to the implications drawn from a review of the research literature.[25] In brief outline, the plan of the investigation consisted of these steps:

1. At the beginning of the semester, the following pitch discrimination tests were given to 16 Northwestern University students: (a) two administrations of the Seashore Pitch Discrimination Test, Series B (standard frequency, 500 ~); (b) two administrations of the Wyatt Pitch Discrimination Test (standard frequency, 465.2 ~); (c) tests with an oscillator at standard frequencies of 250, 500 and 1000 ~.

2. The testing was followed by approximately twelve 50-minute periods of individual training in both pitch intonation and pitch discrimination. With minor exceptions, this training was given at a standard frequency of 500 ~. Attention was centered upon diagnosis of individual difficulties and upon development of the best possible "work methods" for each *S*.

3. Post-training retests, identical with the pre-training tests were given at the end of the semester.

[25] *Vide*, p. 28-29.

SUBJECTS

Since it was not feasible to give intensive training to a large number of cases, it seemed preferable to use a small number of *S*s whose training could be adjusted and individualized in accordance with their particular difficulties. The decision was further influenced by the fact that the *t* test, a relatively new technique for determining statistical significance when the sample is small (8, 10), could appropriately be applied to the data. Moreover, it was believed that if statistically significant results were obtained with a small number of cases, corroboration by repetition with a larger group could follow in normal course.

Sixteen Northwestern University students, ten women and six men, were selected as *S*s. Seven of the students were enrolled in the School of Music and another, U.M., was a professional piano teacher who is classified with the music *S*s even though she happened to be enrolled in the College of Liberal Arts at the time. These eight *S*s are described as the "music group," while the other eight *S*s constitute the "non-music group."

Although the music *S*s were all fairly proficient in singing, practically all of the non-music *S*s had experienced difficulties with pitch. B.H. reported that he had been excluded from grade school music classes because he could not carry a tune and that his singing was still "the point of many jokes;" D.L. said that he could not sing a melody if the person next to him was singing a harmony part; E.V., who had been told by a grade school teacher that she was a monotone and would never be able to carry a tune, wrote in reporting this, "I bowed my

head in childish embarrassment, never to sing again. . . . Even to this day I have not tried. . . . Music means rhythm but has absolutely no melody or tone quality to me;" G.K. stated that she had never been able to carry a tune or stay on key and that she "always got lost in jumping from one note to another;" S.J. reported that his singing was "pretty bad" and that he was unable to carry a tune; never able to get on key, S.Jo. went through music classes merely moving his lips but never actually singing (he referred to himself as a "sparrow"); S.M. unwittingly modulated several times in trying to sing a simple tune written in one key.

Most of the *Ss* were low or mediocre initially in the Seashore and Wyatt tests and they did not improve significantly in pre-training retests. There was no reason to suspect that they were "cognitively" limited, however. They were all college students who had successfully passed stringent entrance requirements. None appeared to be handicapped by inability to understand instructions, poor application or low motivation. Moreover, in the case of at least half of the *Ss*, the music group, a preliminary understanding of the meaning of pitch and pitch differences may be assumed.

The music *Ss* were motivated largely by a desire for self-betterment and at least two of the non-music *Ss*, who had voluntarily asked for permission to participate, were similarly motivated. The cooperation of the other six non-music *Ss* was promoted by excusing them from certain requirements in one of their classes.

TESTING PROCEDURES

The following criteria were used to measure the effects of training: (1) scores and ranks in the revised Seashore Pitch Discrimination Test, Series B; (2) scores and ranks in the Wyatt Pitch Discrimination Test; (3) percentage of error in tests given with an oscillator at standard frequencies of 250 ~, 500 ~ and 1000 ~. The tests and the apparatus used in the testing require brief description.

Seashore Pitch Discrimination Test, Series B

Description. This test is recorded on one face of a 12-inch record as one of six tests which comprise the 1939 revision of the Seashore Measures (27, 28). The source of the stimuli used in the recording was a beat-frequency oscillator with an attached incremental frequency condenser. This apparatus produced "essentially pure" tones with their duration controlled by a tape-timing device and with the intensity held constant. The test consists of 50 pairs of tones with the second tone of each pair either higher or lower than the first. The score is the number of correct judgments. The standard tone is 500 ~ and increments of 8 ~, 5 ~, 3 ~, 2 ~ and 1 ~ are presented in order of difficulty, ten trials at each level. These increments represent a narrower range than is found in the original Seashore pitch test and a more difficult as well as a more restricted range than is found in the revised test, Series A. The B series is intended for musical groups and "selected individuals" and is recommended for use in musical organizations, the music studio and the psychological laboratory.

Standardization. The Series B pitch test was standardized on 752 adults. Norms for the test have been expressed in ranks ranging from 1 to 10, a rank of 1 representing the highest 10% of the population and the rank of 10 the lowest 10%. These ranks are further convertible into ratings with a rank of 1 de-

scribed as "superior," a rank of 2 as "excellent," ranks of 3 and 4 as "good," 5 and 6 as "average," 7 and 8 as "low average" and 9 and 10 as "poor."

Reliability. The corrected split-half reliability coefficient for the Series B pitch test is stated by the authors (28) to be .78 ± .02. Test-retest reliability has not been reported.

The writer gave the Series B pitch test twice (24-hour interval) to two relatively homogeneous groups of music majors at Northwestern University. Corrected split-half *r*s in the first testing were .61 ± .04 and .59 ± .04 (n = 116 and 138). For the first group, mean scores in the two testings were 41.55 and 42.10 with σs of 3.88 and 3.80. For the second group, mean scores were 40.53 and 41.10 with σs of 3.88 and 3.80. Test-retest *r*s for the two groups were .35 ± .06 and .31 ± .05 (n = 112 and 133).

With respect to the relative meaningfulness of split-half *vs.* test-retest methods of computing reliability, Farnsworth (6, p. 302) has concluded in favor of the former, at least in connection with the 1919 Measures. His reasons are as follows: (1) it is difficult to maintain interest in the retest; (2) cues may be noticed in the retest which were not noticed in the first test; (3) there may be a memory carry-over. Hazel Stanton (35, p. 32) has observed that in interpreting correlation values for the Seashore tests, we have been dependent upon interpretations of correlations as given for intelligence tests and other paper and pencil tests, but that there is no published interpretation of correlation for measurements in which the threshold of auditory discrimination is involved. She believes that correlations for these measures should be interpreted with some consideration of this factor.

An additional complication, observed earlier in the paper (p. 9), might be that in tests of this type, the lowest *S*s frequently improve their scores in the retest, while the initially high *S*s often score lower. This tendency might produce fairly stable averages, but it would also contribute toward reducing the test-retest reliability.

Validity. Investigators who have used correlation technique for estimating validity have generally reported low *r*s for the Seashore pitch test (13, 14). Seashore has repudiated most of these studies and has relied heavily upon the findings of Hazel Stanton (32), who found the tests useful in a practical way at the Eastman School of Music. Most of Seashore's observations concerning the validity of his tests of pitch discrimination are in the nature of an appeal to logic, however. The argument runs as follows:

1. Pitch is essential to adequate musical hearing because as the psychological correlates of one of the characteristics of the sound wave it is one of the four fundamental media through which music can be heard and performed.

2. The pitch discrimination tests measure what they purport to measure because timbre, duration and loudness are kept constant and measured deviation in frequency is the only factor which varies.

3. Good pitch discrimination is needed in order to hear and employ artistic deviations from true pitch.

4. Pitch is a "specific" and must be validated as such, i.e., against the role that pitch *per se* plays in the musical situation. Pitch discrimination tests should not be validated against the *total* or "omnibus" musical situation because too many other factors are involved.

5. Good pitch discrimination is not by itself predictive of musical success. It is one of many factors.

6. When properly established, a low rating in pitch discrimination is significant and may be taken as predictive of corresponding difficulties in musical pursuits.

7. Pitch discrimination is basic to a cluster of "variants and compounds," sensory, imaginal, affective and motor.

Procedure in administering the test. In the case of eleven of the *Ss*, both the first test and the retest were administered individually, but five of the non-music *Ss* were first tested along with other

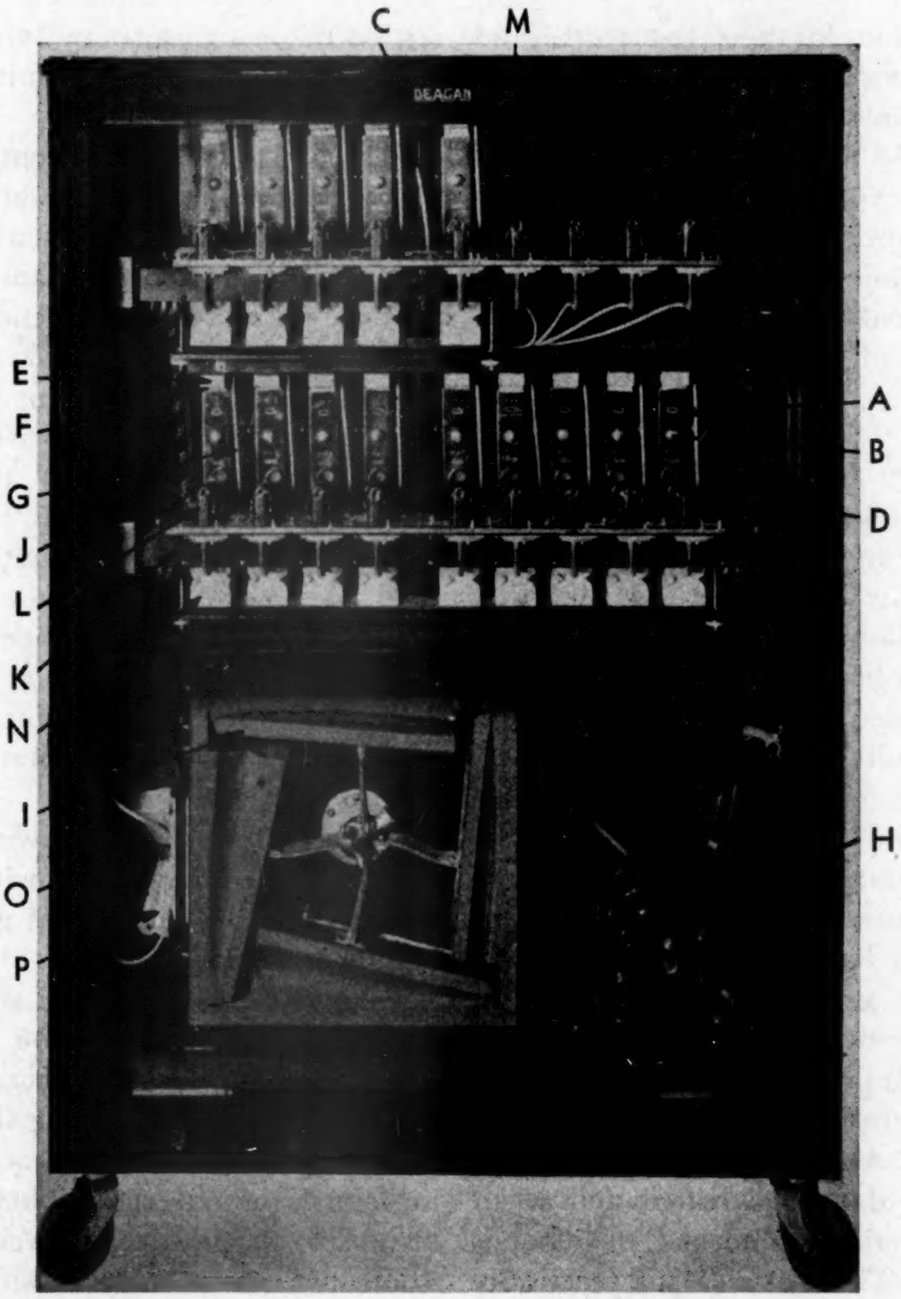

FIG. 1. The front of the instrument used in giving the Wyatt Test of Pitch Discrimination. (The front panels have been removed.) A, steel tuning bar; B, rubber-insulated pin; C, resonator; D, mallet; E, vacuum damper or bellows; F, damper felt; G, damper vacuum line; H, motor; I, quadruple vacuum pump; J, vacuum line; K, vacuum action rail or header; L, electro-magnet for tonal stimulus; M, electro-magnet for intensity test; N, vacuum action for mallet; O, vacuum pressure regulator valve; P, pneumatic.

members of a small psychology class, then retested individually. None of the *Ss* had any opportunity to hear the test during the training period. Post-training tests and retests were individually administered at the end of the semester. Each test was preceded by a fairly lengthy period of demonstration and oral practice.

In addition to these test scores and ranks, there were available for six of the music *Ss* scores and ranks in a test and retest taken 6-8 months prior to the beginning of the experiment as a part of the entrance test program of the School of Music. Thus, before training was begun, ten of the *Ss* had taken the Seashore pitch test twice and six of the *Ss* had taken it four times.

Wyatt Pitch Discrimination Test

The development of a new test of pitch discrimination was begun in 1932. At this time both the Kwalwasser-Dykema Pitch Discrimination Test and the 1919 form of the Seashore Sense of Pitch Test seemed inadequate for precise measurement of adults majoring in music.[26] The test and the instrument described below have been in use since 1933.

Description. The front of the apparatus,[27] illustrated in Fig. 1, shows 14 steel tuning bars (A) mounted on resonators and struck with controlled intensity by felt-tipped rubber mallets (D). Dampers (E) provided at the top of each bar operate in synchronism with the mallet action. At the instant the mallet strikes, the damper is lifted and when the action returns to normal, the damper is released. The motor (H) is practically inaudible, being doubly mounted in rubber. The fact that the vacuum pump (I) is quadruple makes for smoothness of operation. From the main chamber of this pump, vacuum lines (J) extend to the vacuum action rails (K) which carry the electro-magnets (L and M) and serve the two banks of vacuum actions (N). A valve (O) permits regulation of the vacuum within a range of 15-60 inches of water gauge vacuum. The pneumatic action not only insures uniformity in the force of the blow, but permits variation in intensity.

The back of the instrument illustrated in Fig. 2, contains two manual keyboards, one used in giving an intensity test (I) the other (A) used for manual demonstration of different frequencies. In the actual test and also in the written practice exercise, the instrument is operated automatically. One of the perforated rolls (M or E) is placed over the contact roller (D) and as it revolves, the perforations permit small wire contacts (C) to touch the roller. This closes the appropriate circuits. In actual operation, the instrument is closed and *E* merely presses a button on the outside of the instrument in order to start rotation of the roll.

The perforated rolls were cut in accordance with the following factors: the number of perforations governs the number of trials (the rolls used in this experiment each contained slots for 220 stimuli); the lateral position of the perforation determines the bar which is to be struck; the length of the slot controls the duration of the tone (each tone was sustained for one second before it was damped); the distance between the slots controls the time interval between tones (an interval of one second was allowed between the two tones of each trial and three seconds were allowed for the response); perforations appropriately inserted bring the instrument to an auto-

[26] Specific reasons for the inadequacy of these tests are enumerated in other papers (43, 44).

[27] Built by J. C. Deagan, Inc., Chicago, Illinois.

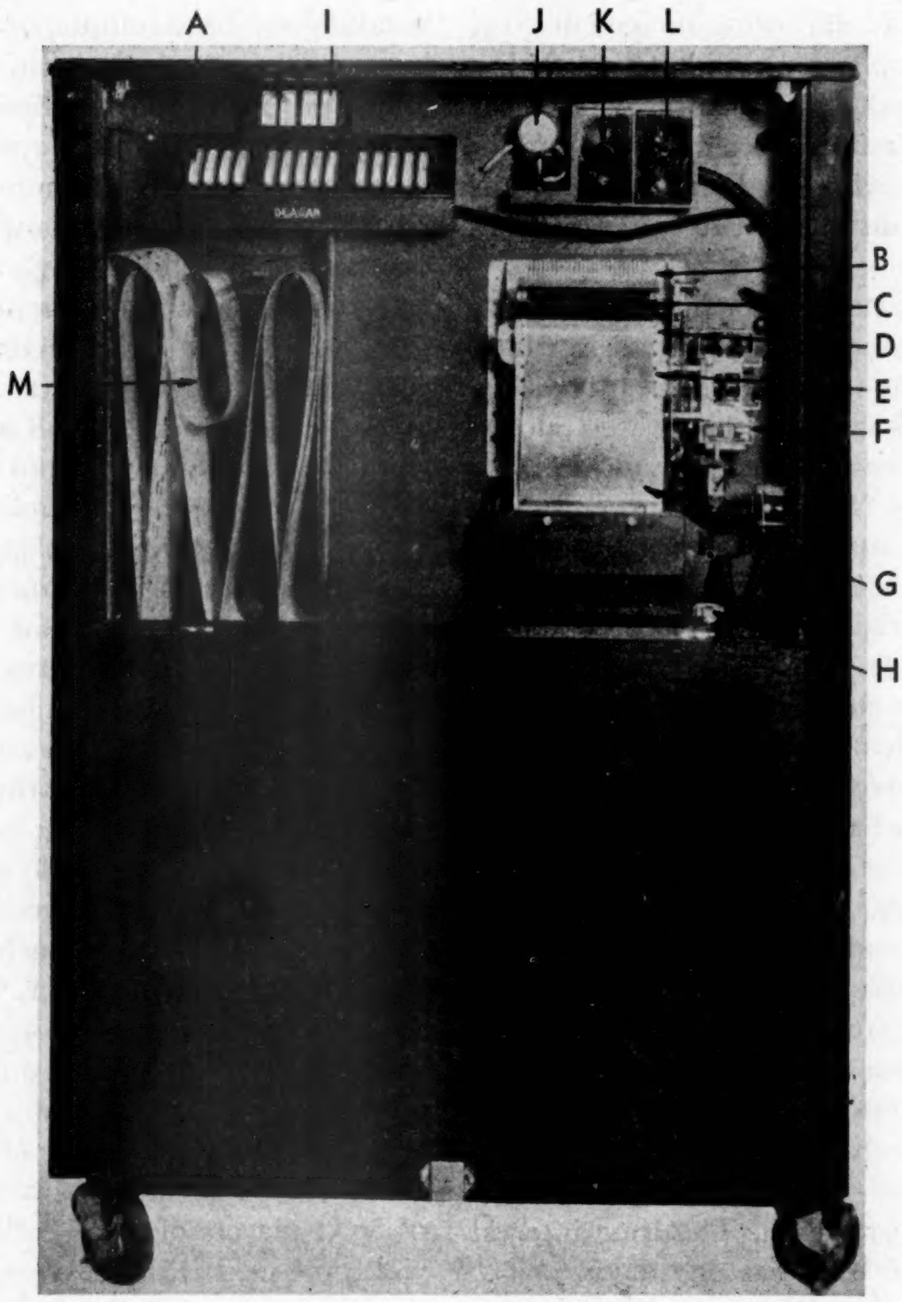

FIG. 2. The back of the instrument used in giving the Wyatt Test of Pitch Discrimination. A, manual keyboard for the tonal stimuli; B, electric player device; C, wire contacts; D, contact roller; E, perforated roll; F, player motor; G, metal roll guard; H, hand adjustment for vacuum pressure regulator valve; I, manual keyboard for intensity test; J, water vacuum gauge; K, pilot light; L, switch; M, extra perforated rolls in storage compartment.

matic stop (a brief pause was allowed after every 20 trials).

The standard bar has a frequency of 465.2 ~.[28] This frequency was chosen arbitrarily, the only specifications imposed being that the frequency be intermediate and that it be some tone other than "violin A" (in order to avoid the possibility that musically trained Ss might have well-developed pitch memory for this particular tone). The comparison bars are higher by 18, 13, 9, 7, 6, 5, 4, 3, 2.5, 2, 1.5, 1 and 0.5 ~ respectively. The test administered in this experiment has two equally difficult forms with ten trials

[28] By cutting a different perforated roll, the standard can be changed to a slightly higher frequency.

at each of the following increments: 13, 7, 5, 4, 3, 2.5, 2, 1.5, 1 and 0.5 ~ making a total of 100 trials. In form A the most difficult increments occur in the middle of the test, while in Form B, every ten trials presents the entire range of difficulties.

A written practice exercise of ten trials follows a demonstration period in which *E* plays the manual keyboard (A in Fig. 2) and *Ss* recite the answers aloud. As in the Seashore test, the form of the response is "higher" or "lower." The score is the number of correct responses, maximum 100.

Standardization. Two sets of norms are available, based on scores of 913 adult music students and 252 adult non-music students respectively. To facilitate comparison with the Seashore test, these ranks have also been expressed as deciles.

Reliability. The correlation of odd-numbered and even-numbered items in the first test taken by 913 music majors was .83 ± .01 (corrected). Considering the homogeneity of the group, this seems adequate. For 213 of these cases, the test-retest reliability was .43 ± .04.[29] Mean scores were 89.1 and 91.5 with S.D.'s of 7.5 and 7.3. The time interval between the tests was one day.

Validity. The Wyatt test has not been used for guidance and was not intended as a "capacity" test. It was designed as an ability test which could be used diagnostically. In order to determine the validity of the test, the writer administered it to 33 musicians who were members of the orchestra and soloists at the National Broadcasting Company. The tests were given in the sound-proofed studios in which the musicians rehearsed and broadcast. Motivation was good with *Ss* taking the tests voluntarily and evincing considerable interest in the apparatus and in their scores. Because of the time factor, the preliminary practice had to be reduced to a minimum and as a result, a few of the tests show some indications of unreliability, e.g., an error at 13 ~ or 7 ~ when no errors occurred at 5, 4, 3, or 2.5 ~. Despite this fact, the group showed negligible error at 13, 7, 5 and 4 ~ (less than 1%) and a very small amount of error at 3 ~ and 2.5 ~ (less than 5%). At 2, 1.5, 1 and 0.5 ~, the error was approximately 9%, 19%, 30% and 34% for these four increments respectively. The threshold for the 33 musicians as a group was between 1.0 and 1.5 ~ and increments larger than 3 ~ were apparently so easy as to be practically non-functional. In this group of professional musicians, six were violinists; four played viola, 'cello or bass viol; 13 played wind instruments; there were four pianists and three singers, a harpist, guitarist and choral director. The harpist and guitarist had scores of 94 and 93; the violinists had a mean score of 92.5; the singers scored 90.7 on the average; the choral director had a score of 90; players of wind instruments had a mean score of 89.5; players of viola, 'cello and bass viol came next with an average 88.8 and the pianists were the lowest in the group with an average score of 85.3. In the entire group, 22 of the 33 had scores which were better than the median of music majors at Northwestern University. Even though a longer time allowance for practice might have improved the scores, those who were required in their work to "make their own pitch" actually did have higher mean scores than the pianists, who had the least need for great precision in pitch discrimination or intonation.

Correlation of Seashore and Wyatt pitch tests. For 227 music majors, the

[29] For interpretation of the test-retest *r* for pitch discrimination tests, *vide* p. 32 this paper.

correlation between scores in the first administration of the Seashore Series B Pitch Discrimination Test and the Wyatt test was .33 ± .04. Several factors may be responsible for the relatively low correlation; (1) the fact that the Seashore test covers a range of increments from 8 ~ to 1 ~, while the Wyatt test covers a range from 13 ~ to 0.5 ~ with 50% of the test at increments smaller than 3 ~; (2) the Seashore test is half as long as the Wyatt test; (3) the duration of the tones and the time interval between tones is not the same in the two tests; (4) the tonal stimuli in the Seashore test are recorded oscillator tones, while the Wyatt test is given by means of tuning bars; (5) the Wyatt test was preceded by both oral and written practice, but only the former type was given for the Seashore test.

Procedure in administering the test. As in the case of the Seashore test, ten of the *S*s had two pre-training administrations of the Wyatt test, while six of the *S*s had four testings, the first two 6-18 months prior to the beginning of the experiment. In all cases the last pre-training test was administered individually. In eleven cases, the first pre-training test that was given as a part of the experimental plan was also individually administered. All post-training tests were given individually.

Each test was preceded by a period of demonstration and oral practice in which the manual keyboard was employed, and, in addition, ten written practice trials, in which the instrument operated automatically as in the actual test, were given.

Tests with the oscillator

Description of the oscillator. A resistance-tuned oscillator[30] was used for both testing and training. This instrument has a frequency range from 20 ~ to 20,000 ~ and a 50 db. range in intensity. An automatic timing device made possible the presentation of paired stimuli with the standard tone occurring either first or second and with automatic control of the duration of each tone (1 second), of the time interval between the standard and the incremental tone (1 second) and of the time allowed for the response (4½ seconds). Switches were clickless and could be moved without presenting any perceptible cue. Two sets of high-fidelity head-phones were used with the oscillator, one worn by *S*, the other by *E*.

Procedure in administering the oscillator tests. *S*s were seated so that they were facing a wall and were unable to see the manipulation of the dials and switches or the recording of the responses. In order to reduce the possibility of adventitious success, 40 consecutive trials were given at each increment. After hearing the two tones of each trial, *S* gave an oral judgment as to whether the second tone was higher or lower than the first. Following one of five keys which had been prepared in advance, *E* recorded the number of errors in the 40 trials presented at each increment. Tests were given individually at three different frequency standards—250 ~, 500 ~ and 1000 ~. The actual increments employed differed for each individual in accordance with his ability. An attempt was made first to determine the smallest increment at which no error, or only a negligible amount of error, occurred in 40 trials. This "initial increment" was not the same for all *S*s. For one individual it was 3 ~, while for another it was 100 ~. The test for each *S* was continued through a progressively more difficult series of increments until approximately

[30] Model 200 S-14, supplied by the Hewlett-Packard Company, Palo Alto, California.

25%-35% error was made or until a difference of 1.5 ~ was presented.[31] This value was the "final increment." The number of increments presented between the initial and final increments varied from as few as three to as many as thirteen, depending upon *S*'s performance. Thus, an oscillator test consisted of as few as 120 trials or as many as 520 trials.

The percentage of error furnished a record of performance in each oscillator test. Since for each *S*, the post-training tests covered exactly the same increments and the same number of trials as the pre-training tests, the percentages of error were comparable.

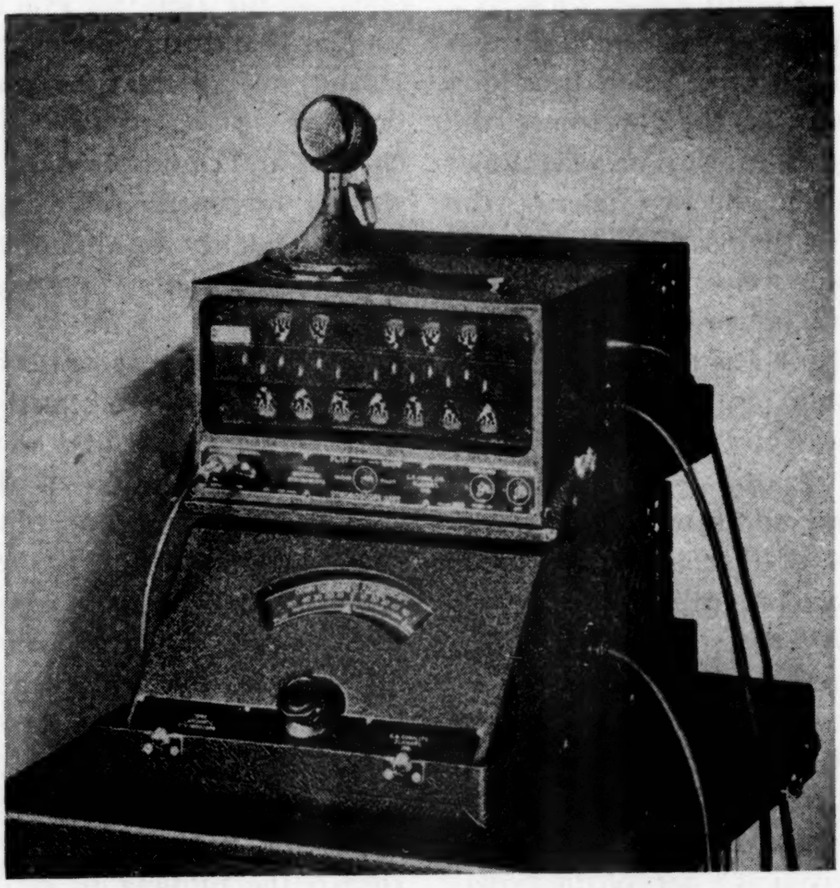

FIG. 3. The Conn Chromatic Stroboscope

TRAINING PROCEDURES

Ss were trained individually in both pitch intonation and pitch discrimination. It was intended that all of the training be given at a frequency of 500 ~, but in a few instances *Ss* could not sing this tone with comfort and for these *Ss* a lower standard had to be employed for the training in intonation. All of the training in discrimination was given at a standard of 500 ~ however. An attempt was made to adjust the method and the type of training to individual needs. For some *Ss*, pitch intonation was stressed, but for others who quickly demonstrated that they were accurate in pitch intonation, effort was concentrated upon de-

[31] Early in the course of the training, but after the pre-training tests had been given, it was discovered that there was a discrepancy of 1 ~ between the dial reading for the increments and their actual frequency. This accounts for the fact that the smallest increment used was 1.5 ~ instead of 0.5 ~ as intended. All results on the oscillator tests are corrected for this discrepancy, which was constant at all frequency levels.

veloping better work methods in discrimination. Approximately twelve 50-minute periods were devoted to training each *S*.

Training in pitch intonation. A Conn Chromatic Stroboscope (46) was available for the experiment. This instrument provided a rapid visual check on the accuracy of *S*'s ability to match a standard tone and to sing intervals, for when a tone was sung into the microphone, it was possible for *S* to see whether the intonation was accurate or, if there was a deviation from the correct pitch, to see the direction and estimate the amount of the error. The essential features of the apparatus are as follows:

Fig. 3 shows the fork unit (below) and the stroboscope unit (above, with the microphone set upon it). There are 12 windows in the stroboscope unit and each window corresponds to one of the chromatic tones of the octave (equal temperament). The arrangement of the windows is patterned after a piano keyboard, the upper row of windows corresponding to the black keys of the piano. The windows are illuminated when sounds are picked up by the microphone. Each window represents a particular tone over a range of seven octaves. When the pointer of the fork unit is set at zero, a frequency of 440 ~ (or octaves) produces a stationary pattern in the 'A' window. If the tone is sharp, the pattern in this window moves to the right; if flat, the pattern moves toward the left. If one wishes to employ some standard frequency which does not conform to an equally tempered interval of 440 ~, the fork unit can be adjusted so as to bring the stroboscope into synchronism with the source.

Specific procedures employed for the training in pitch intonation were as follows:

1. The functioning of the stroboscope was first demonstrated. It was pointed out that the fork unit of the stroboscope was in exact synchronism with an oscillator tone of 500 ~; that accordingly if a vocal tone of this same frequency were picked up by the microphone, it would produce a stationary pattern in one of the windows; that if the tone were slightly higher than 500 ~, the pattern would drift to the right, while if the tone were slightly lower than 500 ~, the pattern would drift to the left. It was also demonstrated that the windows conformed spatially to the semitones of one octave, hence that accurately intoned intervals would also produce stationary patterns.

2. *S*s were instructed to listen to the standard (500 ~) very attentively while it was sounded three times on the oscillator, then to try to sing the same tone into the microphone immediately after the third tone ceased. Each *S* checked his own vocal precision visually and attempted to correct his intonation if the pattern drifted.

3. If *S* was unable to match the standard under these conditions, the procedure was varied in several ways: (a) by using a lower-pitched standard if the 500 ~ tone was beyond *S*'s comfortable vocal range; (b) by asking *S* to sing while the standard was still sounding; (c) by playing the standard tone on the piano instead of the oscillator; (d) by *E's* singing the standard. The latter procedure seemed to facilitate imitation in many instances of inability to match either the oscillator or the piano standard.

4. *S*s were tested and trained in singing whole tones and semitones above and below the standard. When difficulty was encountered, the familiar tunes, "Three Blind Mice" and "By the Sea," were respectively recalled for diatonic and chromatic progressions.

Individual differences in accuracy of pitch intonation were so marked that the

type of training and the time allotted to working with the stroboscope varied greatly. S.Jo., for example, who could not even approximate the standard tone, much less diatonic or chromatic intervals, spent much of the training time in repeated attempts to match a single tone, while U.M., having succeeded with semitones, requested that she be permitted to practice singing still smaller intervals.

Training in pitch discrimination. Remedial training in discrimination was given with the oscillator at a standard frequency of 500 ~. An effort was made to ascertain the nature of each *S*'s particular difficulties and to aid in the establishment of better work methods. The procedures used in this portion of the training were as follows:

1. *Demonstration with foreknowledge of the correct answer.* An increment was chosen which was slightly smaller than the one at which *S* had most recently given 40 correct responses and at which errors were still being made. *S* was instructed to listen attentively, knowing in advance that the second tone would be higher in every case. After listening to a long series of such pairs, the procedure was reversed and a similar series of paired tones was played at the same increment, but with the second tone lower. This was followed by the playing of alternate "highers" and "lowers," still with advance knowledge of what was to come. The demonstration was followed by a 40-trial practice test at the same increment. If errors were made, demonstration was resumed or some other method was tried. If *S* made a perfect record on 40 trials, however, the practice tests continued at more difficult levels.

2. *Informing S as to the correctness of his response.* This technique was used in connection with the method described above. If *S* made an error in a practice test, he was apprised of it and the test was interrupted for a period of demonstration.

3. *Repetition.* When *S* approached his threshold, he was permitted to hear the pairs of tones twice before giving his response. This allowed him to make a comparison, not only of the first and second tones of a pair, but also of the second tone of one pair and the first tone of the next.

4. *Singing methods.* When *Ss* were unable to discriminate large increments, whole-tone or semitone intervals were played on the oscillator and *Ss* were asked to sing with the tones. Although *Ss* were often grossly inexact in vocal reproduction, in this portion of the training no emphasis was placed upon accurate matching of the tones so long as the direction of the difference was correct. These singing methods were used in combination with 1 above and later with practice tests, *S* giving his response after attempting to imitate the two tones.

5. *Verification of pitch differences with the stroboscope.* The stroboscope was adjusted so that a stationary pattern would appear when the standard tone of 500 ~ was sounded into the microphone. When *S* had difficulty with an increment and claimed that the two tones sounded alike, *E*'s headphones were held to the microphone of the stroboscope so that *S* could simultaneously (1) listen to the pairs of tones; (2) watch the alternate stationary and moving patterns in the appropriate window; (3) respond "higher" or "lower," the former if the pattern for the first tone was stationary while that for the *second* tone drifted to the right, the latter if the pattern for the *first* tone drifted to the right while that for the second tone remained stationary.

6. *Analogy with the diatonic scale.* For

Ss who could not discriminate large increments, the suggestion was made that they think of the two tones as a part of a larger whole. A diatonic scale was played on the piano and S was encouraged to attend especially to the beginning of a descending scale and the ending of an ascending scale ("do-ti" or "ti-do"). Semitone sequences were then played on the oscillator or on a piano and S was asked to tell whether the scale was "on its way up" or "on its way down." Singing methods were often combined with this procedure.

7. *Correction of constant error.* Analysis was made of each S's errors in the pretraining tests and in the practice tests to ascertain whether there was a striking preponderance in one direction. In such cases, the data were shown to S and corrective work was begun.

8. *Recognition tests and anticipatory judging.* The following four-step process was used: (a) Auditory imagery was defined and discussed and some time was devoted to having S listen to the standard tone and then, during an interval of silence, try to form a clear auditory image of the tone. (b) A relatively easy recognition test was then given in which the standard tone was presented along with other tones which were considerably higher in pitch and S was asked to judge whether a tone was the standard or some other tone, i.e., the response took the form of "that is the standard" or "that is not the standard." (c) If S succeeded in recognizing the standard, he was then told that he would hear pairs of tones; that the standard tone would occur in every pair; that all other tones used would be higher in pitch; that if he could recognize the standard, he could give his judgment before hearing the second tone, for if the first tone *was* the standard, the second tone would be higher in pitch, while if the first tone was *not* the standard, the second tone would be the standard and it would be lower in pitch. A practice test was then given in which an anticipatory response, made before hearing the second tone, took the form of "the second tone will be higher" or "the second tone will be lower." Relatively easy increments were used at this stage. (d) Following this, S was asked to anticipate the direction of the second tone as before, but to do so silently and tentatively, then to verify this tentative judgment after hearing the second tone, when the response could be given orally. If this method proved helpful, progressively smaller increments were tried.

IMPROVEMENT OF WORK METHODS

Throughout the training period and again at the end of the semester, verbal reports were elicited from Ss regarding "work methods" in discriminating pitch. It was thought that a study of these comments of Ss, especially when coupled with observation of Ss by E and considered in relation to changes in performance, might lead to a better understanding of what constitutes good work method in pitch discrimination.

Analysis of these factors indicated that in general the most helpful training procedures were those which encouraged the use of auditory imagery and motor participation. It was also observed that postural attitudes were significant, but that there were marked individual differences, so that no one postural attitude was optimal for all. These conclusions can best be illustrated by reference to specific Ss.

Auditory imagery. Inasmuch as pitch discrimination involves comparison of tones which are sounded in succession with a brief time interval between them, an auditory image of the first tone must

be retained long enough to be present with some degree of vividness when the second tone is sounded. Training therefore included procedures which, it was believed, fostered the use of latent ability to retain a clear auditory image and which entailed sufficient repetition to make the use of auditory imagery habitual. In this connection, the records and comments of E.V., S.J., G.M. and U.M. merit special consideration.

E.V. was exceedingly poor initially in pitch discrimination. In the oscillator test at a standard of 500 ~, 4.0%, 7.5% and 10.0% error was made at such large increments as 70 ~, 60 ~ and 50 ~ respectively, while in the test at 1000 ~, 16.7% error occurred at an increment of 100 ~. Because of this initial deficiency, the pre-training tests with the oscillator extended over an unusually long period (almost four hours). During the course of this lengthy pre-training testing, the percentage of error suddenly decreased as the increments became more difficult. Questioned as to whether she had altered her method in some way, E.V. replied: "Yes, I am trying to remember the first tone better; I keep it inside of myself until I get the second tone." In this instance it was spontaneously realized, even before the formal training was begun, that imagery was important and the work method was revised accordingly.

After practice in anticipatory judging, as described in the preceding section, S.J. reported that he used the following variation of the method: As soon as he heard the first tone, he tried to form an image of a higher tone. If, when the second tone was presented, it seemed to "agree" with his image, he responded "higher," but if the tone seemed to "contradict" his image, he responded "lower." Improvement in the practice tests was noted following the adoption of this work method. Similar observations were made by G.M.: "After I hear a few trials, I get the vibrations in my ear . . . I anticipate the second tone. If I think it will be lower, I put my chin down and if I don't feel contradicted, I answer 'lower'."

In a 20-trial practice test with the oscillator set at a standard of 500 ~, U.M. had 30% error on the 2 ~ increment. When asked to give an anticipatory judgment before the second tone was heard, the magnitude of error was reduced to 15%. It was further reduced to 10% when the anticipatory judgment was made silently and the oral response was delayed until after the second tone was actually heard. After a brief interval during which she listened to ten pairs of tones with foreknowledge of the correct answers, U.M. took another practice test and made a perfect score. Thus, in one half-hour period, U.M.'s record in discrimination of a 2 ~ increment changed from 70% correct (the approximate threshold) to 100% correct. In a verbal report, U.M. stated.

A difference of four or five cycles, which formerly gave me uncertain moments and sounded like one tone, now seems very far apart and it is no effort to distinguish the direction of the pitch difference. . . . For those pitch differences that have been mastered, there was guessing at first, then . . . came a feeling for the difference and finally the difference was so apparent that it was difficult to realize that the interval had ever given me a moment of doubt.

Motor participation. Analysis of the verbal reports or actual observation by *E* indicated that in the case of at least five of the *Ss*, kinesthetic sensation and imagery accompanied pitch discrimination. B.H. stated that he 'sort of whispered the tones to himself and tried to determine the differences by corresponding changes in his throat muscles.' Similar comments were made by D.L., G.K., I.D. and T.J. Even when the differences were as small as 1 or 2 ~, I.D. reported that she was imitating the tones subvocally, associating them either with "do-ti" or "ti-do." Occasionally T.J. could actually be heard attempting to imitate the tones vocally before giving his judgment. The intonation was not exact, but the method seemed helpful nevertheless. It is possible that imitative singing in the early stages of training gradually becomes implicit as *Ss* become habituated to this method. In any case, training procedures which encouraged imitation, either vocal

or sub-vocal, seemed to facilitate discrimination in many cases.

Individual differences in postural attitudes. Seashore has suggested that testers direct listeners to "take a position of muscular tension, leaning forward with muscles firm in the most favorable position for writing, in an attitude of attention, eyes closed while listening" (23, p. 6). In view of this recommendation, it seemed quite surprising that two of the *Ss*—I.D. and F.M.—had a larger percentage of correct judgments in the practice tests when they seemed most relaxed and, indeed, almost indifferent (examining their manicures, toying with their hair, looking about the room, etc.) than when they were being formally tested on the same increments and assumed an attitude of strict attention.[32] Individual differences were also observed with respect to the advisability of keeping the eyes closed while listening to the tones. Some of the *Ss* said that they found this helpful and could even be seen placing their hand over their eyes, but others found it better to keep their eyes open. P.E., for example, reported that she was distracted when she kept her eyes closed because she "drew pictures." She preferred to fixate a particular place on the wall while listening. It would appear, therefore, that caution should be used in advising *Ss* as to the optimal postural attitude to assume.

[32] As early as 1914, one of Smith's *Ss* and Smith himself observed that in their own cases, more accurate judgments were made in a state of relaxation than in an attitude of strict attention (cf. p. 19).

IV. Results

THE FOLLOWING quantitative data are available:

1. Scores and ranks in 2-4 pre-training and 2-4 post-training administrations of the Seashore Series B pitch test.[33]

2. Scores and ranks in 2-4 pre-training and 2-3 post-training administrations of the Wyatt pitch test.[34]

3. Pre-training and post-training percentages of error in oscillator tests (the magnitude of the increments depending upon individual performance) at standards of 250 ~, 500 ~ and 1000 ~.

PERCENTAGE OF IMPROVEMENT

a. Seashore pitch test. For each of the sixteen *Ss*, the mean score for the aggregate of all of the pre-training tests taken[35] may be compared with the mean score in the aggregate of all post-training tests taken.[36] A comparison of these pre-training and post-training means is shown in Table 5.[37]

It may be observed that nine of the sixteen *Ss* were substantially higher in their post-training performance (6.5 to 14 points), that six of the *Ss* had small increases (0.5 to 3.5 points) and that one *S* actually scored a little lower after training than before.

The average increase for the entire group was approximately 6 points, with an average gain of 7.75 points for the music *Ss* and of 4.50 points for the non-music *Ss*.

These post-training increases may also be expressed in terms of the maximum possible increase. In relation to their pre-training performance, the post-training increase for the entire group was about 37% of the maximum possible increase, i.e., the improvement necessary to attain the maximum score of 50.[38] The music group made up about half (49%) of the difference between their pre-training mean and a score of 50, while the non-music group came about one-fourth of the way (26%).

TABLE 5
Mean pre-training and post-training scores in the Seashore pitch test, series B

	Ss	Pre-tr. Mean	Post-tr. Mean
Music	DA	33.8	42.5
	FM	29.0	43.0
	GM	34.0	45.5
	ID	36.5	43.0
	PE	37.0	40.0
	RD	44.0	41.5
	SL	29.3	37.0
	UM	30.5	43.5
	Average	34.25	42.00
Non-music	BH	36.5	38.0
	DL	31.0	41.5
	EV	31.5	32.0
	GK	31.5	32.5
	SM	34.0	35.5
	SJ	32.5	42.0
	SJo	30.5	38.5
	TJ	36.5	40.0
	Average	33.00	37.50

The above figures were obtained by comparing aggregate averages in all of the pre-training tests taken (2-4 for each *S*) with aggregate averages in all of the post-training tests taken (2-4 for each *S*). It was considered possible, however, that

[33] *Vide*, p. 33-34.
[34] *Vide*, p. 37.
[35] Ten of the *Ss* had two pre-training tests and six of the *Ss* had four such tests, the first two taken 6-18 months before the experiment began.
[36] Three of the *Ss* were given three post-training tests and two of them received four such tests. All others had two post-training tests.
[37] Tables 5-8 and Figures 4 and 5 present averages, but the original data for each *S* in each test taken may be found in the writer's doctoral dissertation (46, pp. 160-210).
[38] The pre-training mean for the entire group was 33.625. This is 16.375 points under the maximum possible score. The post-training mean of 39.750 made up 6.125 points of this difference. The percentage of improvement may be expressed, therefore, as 6.125/16.375, or 37.4%.

some of the gain might be attributable to an unreliably low score in the first pre-training test, especially since this test was not individually administered in all cases. This possibility was explored by using as a base the scores in the last pre-training test taken (given individually to all *Ss*) and by computing the differences between these scores and scores in the last post-training test taken. The results were found to be very similar, however, to the results obtained when the averages of all of the pre-training and all of the post-training tests were used. The mean gains were 7.87 and 5.88 points for the music and non-music groups respectively (as compared with mean gains of 7.75 and 4.50). The percentage of improvement, i.e., the degree to which these obtained gains approached the total possible gain, was about the same for the music *Ss* (48%) and a little larger for the non-music *Ss* (34%).[39]

b. Wyatt pitch test. Most of the data shown in Table 6 were derived by averaging scores in two pre-training tests and two post-training tests. A few of the *Ss* had additional tests, however.[40]

All sixteen of the *Ss* had post-training gains with a mean increase of 13.60 points for the group as a whole and mean increases of 12.25 and 14.95 points for the music and non-music groups respectively. For both groups, these increases represent approximately 56% and 47% of the increase necessary to attain the maximum possible score of 100.[41]

As in the case of the Seashore test, a further check was made to see whether post-training increases might be attributable, in part at least, to unreliably low scores in the first pre-training test. A comparison of scores in the last pre-training and the last post-training tests taken disclosed mean increases which were very similar in magnitude to the mean increases found when aggregate averages were compared (13.88 and 13.25 for the former comparison; 12.25 and 14.95 for the latter).[42] Percentages of improvement were also similar.[43]

TABLE 6

Mean pre-training and post-training scores in the Wyatt test

	Ss	Pre-tr. Mean	Post-tr. Mean
	DA	77.5	85.5
	FM	76.0	89.0
	GM	68.5	91.0
Music	ID	76.0	86.0
	PE	89.0	96.5
	RD	73.5	89.0
	SL	80.5	93.0
	UM	84.0	93.0
	Average	78.13	90.38
	BH	72.0	85.0
	DL	79.5	92.5
	EV	61.0	73.3
Non-music	GK	70.0	74.0
	SM	75.5	86.0
	SJ	77.0	84.3
	SJo	43.0	85.0
	TJ	69.5	87.0
	Average	68.44	83.39

c. Oscillator tests. Incrementts in the three tests given with the oscillator were

[39] Averages for the last pre-training and the last post-training tests taken by the music *Ss* were 33.75 and 41.62 (as compared with 34.25 and 42.00 for the aggregate averages shown in Table 5). Parallel figures for the non-music *Ss* were 32.62 and 38.50 (as compared with aggregate averages of 33.00 and 37.50).

[40] Two *Ss* had been given one and two tests respectively 6-18 months before the experiment began. Two of the *Ss* were given an additional post-training retest.

[41] In terms of a score of 100, the music *Ss* gained 12.25/21.87, or 56% and the non-music *Ss* gained 14.95/31.56, or 47%

[42] For the music group, the mean scores in the last pre-training and the last post-training tests were 76.75 and 90.63 (as compared with 78.13 and 90.38 for aggregate averages). Parallel figures for the non-music *Ss* were 71.75 and 85.00 (as compared with 68.44 and 83.39 for aggregate averages).

[43] For the music *Ss*, the percentage of improvement was 60%, for the non-music *Ss*, 47%.

selected in accordance with each *S*'s ability. The increments ranged in each case from an "initial increment" (defined as that value at which no errors, or only a negligible amount of error, occurred in at least 40 consecutive trials) to a "final increment" (defined as either the 1.5 ~ increment or that increment at which 25%-35% error occurred in at least 40 consecutive trials). Post-training tests covered exactly the same range of increments as the pre-training tests.

A very marked contrast between the music and non-music groups as to ability and homogeneity was apparent in the oscillator tests. At a standard of 500 ~, for example, the music *S*s had initial increments within a range from 6.0 ~ to 8.5 ~, average 7 ~. The non-music *S*s, however, had initial increments which varied from 8.5 ~ to an increment as great as 100 ~ and the average for the eight non-music *S*s was 28 ~.

Table 7 shows the percentages by which each *S* succeeded, after training,

TABLE 7
Percentages of improvement in the oscillator tests

	*S*s	250 ~	500 ~	1000 ~
Music	DA	65.2	76.4	51.8
	FM	38.2	84.1	39.0
	GM	34.8	77.5	22.3
	ID	6.9	47.5	40.9
	PE	51.5	75.4	32.1
	RD	84.9	83.1	67.9
	SL	32.1	63.0	56.8
	UM	—*	82.4	—*
	Average	44.8	73.7	44.4
Non-music	BH	94.5	94.2	73.5
	DL	81.6	83.0	71.6
	EV	85.4	69.8	80.8
	GK	25.0	51.4	60.9
	SM	30.0	53.4	53.5
	SJ	14.3	93.9	76.5
	SJo	65.7	81.8	40.7
	TJ	11.1	73.5	51.3
	Average	51.0	75.1	63.6

* Having entered the experiment late, UM was not tested at these levels.

in eliminating pre-training errors.[44] All *S*s improved in the discrimination of oscillator tones, with the greatest mean improvement occurring at the frequency level at which practically all of the training was given, viz., 500 ~. At this standard, both groups eliminated about three-fourths of their pre-training error. These percentages also indicate a transfer of improvement to the two other frequency standards. At the 250 ~ and 1000 ~ standards, the music *S*s made only about 45% as many errors as they made prior to training. The non-music *S*s reduced their pre-training error by 51% and 64% at the lower and higher frequency standards. The fact that these percentages are higher for the non-music *S*s suggests that, relative to their rather poor showing in the pre-training tests, they made a somewhat greater advance.

STATISTICAL SIGNIFICANCE OF THE DIFFERENCES

The *t* test was applied in order to determine the statistical significance of the obtained differences in pre-training and post-training performance in the various tests. The formula for deriving *t* (which is essentially a critical ratio for

[44] The record for one individual may serve as an illustration of the method of computing the percentages. At a standard frequency of 500 ~, G. M. was tested at each of the following six increments: 6 ~, 5 ~, 4 ~, 3 ~, 2 ~, and 1.5 ~. At these six increments, the percentages of pre-training error were respectively 0.00, 6.36, 8.75, 18.75, 20.00 and 33.33. Post-training retests employed these same increments, but G.M. made a perfect record in the four largest increments and reduced the error in the two smallest increments to 11.25% and 8.33%. Thus, while the aggregate of the percentages of pre-training error was 87.19, the aggregate of the percentages of post-training error was reduced to 19.58—a difference of 67.61. If G.M. had succeeded in giving 100% correct answers in the post-training test, she would have eliminated 87.19/87.19 of her pre-training error. Inasmuch as 67.61/87.19 was eliminated, the improvement may be expressed as 77.5%.

estimating the significance of a difference between two means) is a highly conservative one, modified to make it particularly suitable for small samples. The smaller the sample, the larger the *t* required for any given level of confidence, e.g., a *t* of 2.947 is significant at the 1% level of confidence for sixteen cases, but for only eight cases, *t* must be 3.499 to be significant at the same level.[45] In an effort to interpret the data in this experiment as conservatively as possible, the eight music *Ss* and the eight non-music *Ss* have been regarded as two independent samples.

a. Seashore pitch test. When the aggregate of all pre-training test scores were compared with the aggregate of all post-training test scores, it was found that the music *Ss* had a mean increase of 7.75 points and that the non-music *Ss* had a mean increase of 4.50 points (Table 5). These gains yield *t* values of 3.993 and 3.067 respectively. The former falls at the 1% level of confidence, the latter at the 2% level.[46]

As a check on the possibility that the first pre-training test scores may have been unreliably low, a further comparison was made between scores in the last pre-training and the last post-training tests. Mean gains of 7.87 and 5.88 points for the two groups respectively yielded *t* values of 3.332 and 3.758, also significant at the 2% and 1% levels of confidence.

b. Wyatt pitch test. Similarly significant *t* values were found for comparisons between pre-training and post-training scores in the Wyatt test. The gains of 12.25 and 14.95 points in the aggregate averages of music and non-music *Ss* gave *t* values of 6.980 and 3.631 respectively. Both are significant at the 1% level of confidence.

For the two groups of *Ss*, *t* values were 4.102 and 2.561 when scores in the last pre-training and the last post-training tests were compared. The former *t* is sig nificant at the 1% level of confidence, the latter at the 5% level.

c. Oscillator tests. Differences in pre-training and post-training percentages of error in the three oscillator tests yielded the following *t* values:

Ss	*Standard Frequencies*		
	250 ~	*500 ~*	*1000 ~*
Music *Ss*	3.883	16.248	7.161
Non-music *Ss*	3.244	11.698	5.025

All of these values are so large that we may be reasonably confident that the observed differences are not due to chance. Five of the *t* values are significant at the 1% level of confidence and one of them falls at the 2% level. As might have been expected, the largest values were found for the 500 ~ test, since it most closely conformed to the examples used in the training. Significant transfer to the other two frequency levels also seems to have taken place.

CHANGES IN RANK

a. Seashore pitch test. Raw scores in the Seashore test are convertible into ranks ranging from 1 to 10. A rank of 1 is interpreted as "superior," a rank of 2 as "excellent," ranks of 3 and 4 as "good," 5 and 6 as "average," 7 and 8 as "low average" and 9 and 10 as "poor." Table 8 shows the ranks which correspond to the pre-training and post-training averages shown in Tables 5 and 6. It may be observed that in their average pre-training performance, only five of the sixteen *Ss* exceeded the median for the

[45] If *t* falls at the 5% level of confidence, it is regarded by Fisher (8) as "significant," while if it falls at the 1% level, it is regarded as "highly significant."

[46] All of the levels of confidence reported here may properly be halved, for we are not considering the significance of differences in *either* direction, but are concerned only with the significance of difference in a *positive* direction (8, p. 126).

population used in standardizing the test. After training, however, thirteen cases exceeded the median.

In their average post-training performance, at least half of the *Ss* had classifications which were significantly higher than their average pre-training classifications. The music group as a whole changed its status from "low average" (a rank of 7) to "excellent" (a rank of 2). The non-music group as a whole changed its status from "low average" (a rank of 7) to "good" (a rank of 3.5).[47]

Similar results were obtained in a comparison of ranks in the last pre-training test and the last post-training test taken. Although in all cases these tests were given individually and although these were retests, thus complying with Seashore's suggestions for overcoming cognitive factors, 75% of the cases greatly changed their pre-training status, as shown below:

Pre-training status	Post-training status	N
Poor	Good	2
	Excellent	1
	Superior	2
Low average	Excellent	3
	Superior	2
Average	Excellent	1
	Superior	1

b. Wyatt pitch test. This test has two sets of norms based on scores of music and non-music University students. To keep the results roughly comparable with those in the Seashore pitch test, ranks are similarly expressed and interpreted.

Prior to training, both groups would have been classified as "poor" in relation to their respective populations. In their post-training performance, however, the two groups improved sufficiently to bring them up to the median performance of the populations on whose scores the norms were based. The post-training rank of each of the sixteen *Ss* was superior to the pre-training rank. Of the fifteen cases who were below the median before training, only three failed to equal or exceed the median after training.

Similar results were found in the comparison of the last pre-training and post-training tests taken. In the pre-training distribution, fourteen *Ss* were below the median for their respective populations, but only five of these cases failed to equal or exceed the median after training.

TABLE 8

Pre-training and post-training ranks in the Seashore and Wyatt pitch tests

	Ss	Seashore test Pre-tr.	Seashore test Post-tr.	Wyatt test* Pre-tr.	Wyatt test* Post-tr.
Music	DA	7.0	1.5	9	4
	FM	9.0	1.0	10	5
	GM	7.0	1.0	10	4
	ID	4.5	1.0	10	7
	PE	4.0	2.0	5	1
	RD	1.0	2.0	10	5
	SL	9.0	4.0	9	3
	UM	8.5	1.0	7	3
Non-music	BH	4.5	3.0	9	4
	DL	8.0	2.0	7	1
	EV	8.0	8.0	10	8
	GK	8.0	7.5	9	8
	SM	7.0	5.5	8	4
	SJ	7.5	2.0	7	5
	SJo	8.5	3.0	10	4
	TJ	4.5	2.0	9	3

* The Wyatt test has different norms for music and non-music *Ss*.

CHANGES IN PERFORMANCE AT VARIOUS LEVELS OF DIFFICULTY

a. Seashore pitch test. The Seashore pitch test has five levels of difficulty—8 ~, 5 ~, 3 ~, 2 ~ and 1 ~ increments—with ten trials at each level. The average percentages of pre-training and post-training error for the entire group of sixteen *Ss* were:

[47] These classifications are not derived by averaging the ranks shown in Table 8, but by converting the mean raw scores shown in Table 6 according to the Manual (27).

	8 ~	5 ~	3 ~	2 ~	1 ~
Pre-tr. error (%)	16.7	28.6	31.3	38.3	49.8
Post-tr. error (%)	00.3	9.4	25.4	26.9	42.1

These data are shown graphically in Fig. 4. As expected, in both the pre-training and post-training tests, the percentage of error became greater as the size of the increment diminished. Relative to pre-training performance, the greatest improvement occurred at the two largest increments.

Prior to training, the music *Ss* were somewhat superior to the non-music *Ss* in discrimination of the two largest increments,[48] but at the three more difficult increments, the pre-training percentages of error for the two groups were very similar. After training, however, the dissimilarity between the two groups was eliminated at the 8 ~ increment, but at all other levels of the test, the music *Ss* were superior to the non-music *Ss*.[49]

As a group, the music *Ss* improved markedly at all levels of difficulty in the test, reducing the magnitude of their error at the five increments respectively by 100%, 75%, 59%, 45% and 21%. The non-music *Ss* as a group improved greatly in discrimination of the two larger increments, but at the three more difficult increments, the changes were relatively small.

b. Wyatt pitch test. This test has ten trials at each of ten increments— 13 ~, 7 ~, 5 ~, 4 ~, 3 ~, 2.5 ~, 2 ~, 1.5 ~, 1 ~ and 0.5 ~. The average pre-training and post-training percentages of error for the group as a whole were as follows:

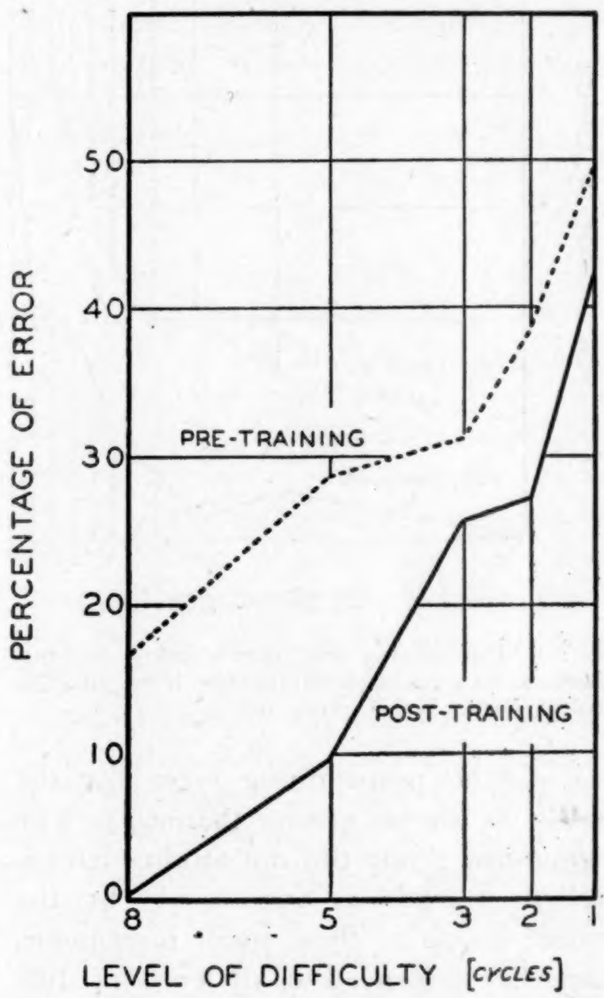

FIG. 4. Pre-training and post-training percentages of error at each of the five levels of difficulty in the Seashore pitch test.

	13 ~	7 ~	5 ~	4 ~	3 ~	2.5 ~	2 ~	1.5 ~	1 ~	0.5 ~
Pre-tr. error (%)	10.8	13.9	20.1	19.7	24.1	30.9	35.3	41.9	39.0	35.0
Post-tr. error (%) ..	00.0	4.0	2.0	6.3	7.9	14.5	18.0	20.9	21.6	38.9

These data are shown graphically in Fig. 5. At all levels excepting the most difficult 0.5 ~ level, the magnitude of error was greatly reduced (by at least 45%). Improvement was indicated by the music

[48] The error for the music *Ss* was about 10% and 23%. For the non-music *Ss*, the error was approximately 23% and 34% at the same levels of difficulty.

[49] Post-training percentages of error for the music *Ss* at the five levels of difficulty were: 0.0, 4.8, 13.1, 23.5 and 38.5. Corresponding figures for the non-music *Ss* were: 0.6, 14.0, 37.7, 30.2, and 45.6.

Ss at all ten levels of difficulty.[50] All errors were eliminated at 13 ~ and 7 ~ and only negligible error was made at 5, 4 and 3 ~ (less than 4%). Even at the relatively difficult increments of 2.5, 2 and

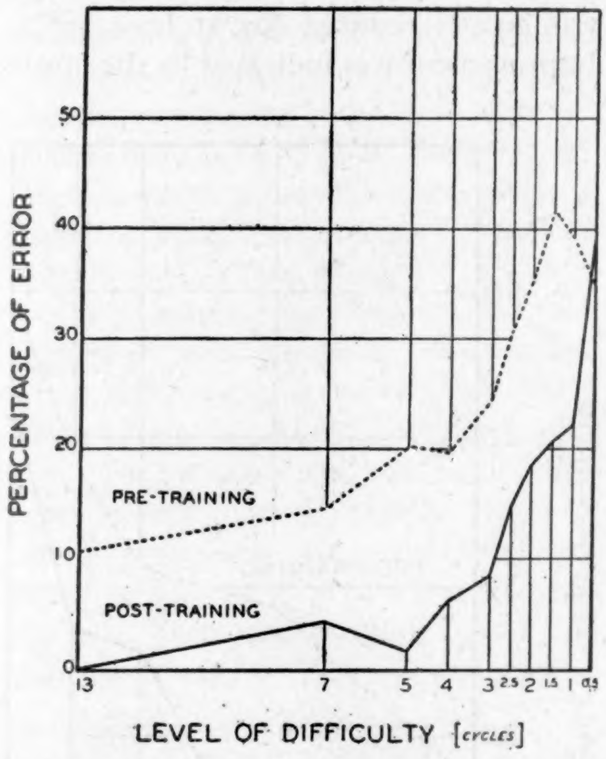

FIG. 5. Pre-training and post-training percentages of error at each of the ten levels of difficulty in the Wyatt pitch test.

1.5 ~, the post-training error for the music *Ss* was no greater than 15%. The non-music group did not attain such excellent records as were made by the music group at these small increments, but they improved at all levels of difficulty excepting the 0.5 ~ increment, eliminating all errors at 13 ~ and reducing their percentage of error to 12% or less at increments from 7 ~ to 3 ~ inclusive.[51]

c. Oscillator tests. Increments in the oscillator tests were selected in accordance with each *S*'s ability, ranging from the "initial increment" (the level at which no errors, or only a negligible amount of error, occurred in at least 40 consecutive trials) to a "final increment" (either the 1.5 ~ increment or the increment at which 25%-35% error was made). The average pre-training and post-training initial increments for the entire group of sixteen *Ss* were as follows:

	250 ~	*500 ~*	*1000 ~*
Pre-tr.	12.4 ~	17.6 ~	28.0 ~
Post-tr.	8.3 ~	9.2 ~	20.0 ~

Each of these values was consistently smaller for the music *Ss* than for the non-music *Ss*. The average pre-training initial increments for the two groups respectively were:

	250 ~	*500 ~*	*1000 ~*
Music *Ss*	4.8 ~	6.9 ~	11.7 ~
Non-music *Ss*	19.0 ~	28.3 ~	42.2 ~

Average post-training initial increments for the two groups were:

	250 ~	*500 ~*	*1000 ~*
Music *Ss*	4.1 ~	4.4 ~	8.0 ~
Non-music *Ss*	11.5 ~	14.0 ~	30.8 ~

It is clear that the difference between pre-training and post-training performance was most pronounced for the initial increments at a frequency standard of 500 ~—the level at which the training was given—but that transfer to the two other frequency levels also occurred.

No effort was made in this study to ascertain exact psychophysical thresholds. *Ss* were tested at progressively more difficult increments, but since it was not practical to employ increments smaller than

[50] For the music *Ss*, reductions in the percentages of error at the ten levels of difficulty were as follows: at 13 ~, from 0.9% to 0.0%; at 7 ~, from 2.1% to 0.0%; at 5 ~, from 8.9% to 0.6%; at 4 ~, from 14.5% to 1.9%; at 3 ~, from 18.9% to 3.7%; at 2.5 ~, from 26.2% to 10.6%; at 2 ~, from 32.4% to 15.0%; at 1 ~, from 38.7% to 13.7%; at 1 ~, from 37.4% to 20.0%; at 0.5 ~, from 38.7% to 36.2%.

[51] For the non-music *Ss*, average percentages of error at the ten levels of difficulty changed as follows: at 13 ~, from 20.6% to 0.0%; at 7 ~, from 25.6% to 7.9%; at 5 ~, from 30.6% to 3.3%; at 4 ~, from 25.0% to 10.6%; at 3 ~, from 29.4% to 12.1%; at 2.5 ~, from 36.6% to 18.3%; at 2 ~, from 38.1% to 21.0%; at 1.5 ~, from 45.0% to 28.1%; at 1 ~, from 40.6% to 23.1%; at 0.5 ~, from 31.3% to 41.5%.

1.5 ~, the actual threshold was not approximated in many instances. There is, however, a well-defined trend toward improvement after training, indicated either by lowering of the approximate threshold, if it had been ascertained, or by reduction in the percentage of errors at the 1.5 ~ increment. For the entire group, the average final increments were as follows:

	250 ~		500 ~		1000 ~	
Pre-tr.	3.30	21%	2.34	24%	3.50	27%
Post-tr.	1.60	12%	1.75	16%	2.63	20%

The pattern of improvement is consistent, again confirming the existence of transfer of the effects of training.

SUMMARY OF RESULTS

a. *Seashore pitch test.* Each of sixteen adult *Ss*—eight music *Ss* and eight non-music *Ss*—was given the Seashore Series B pitch test at least twice before training was begun and at least twice after training was completed. Some of the *Ss* were tested even more intensively.. The sixteen *Ss* had an aggregate of 44 pre-training testings and 39 post-training testings. Two types of comparison were made, based on different sets of data: (1) the average performance in all of the pre-training tests was compared with the average performance in all of the post-training tests, thus utilizing all of the data available; (2) as a check on the possibility that too great weight might be placed upon the first pre-training test, which might be unreliably low, performance in the last pre-training test was compared with performance in the last post-training test taken. Only the results based on aggregate averages (1 above) need be reviewed here, as the results in both types of comparison were so similar:

1. The mean post-training gain for all sixteen *Ss* was 6.125 points. This represents about 37% of the maximum possible gain. The mean increase for the music *Ss* (7.75 points) was greater than that for the non-music *Ss* (4.50 points). The former group achieved approximately 49%, the latter approximately 26% of their maximum possible gain.

2. For the two groups respectively, the differences between pre-training and post-training means yielded *t* values which were "highly significant" at the 1% and 2% levels of confidence.

3. Nine of the sixteen *Ss* improved greatly in their post-training tests, with increases of 6.5-14.0 points over their pre-training scores.

4. The average rank of the music *Ss* changed from 7, interpreted as "low average," to 2, interpreted as "excellent." The average rank of the non-music *Ss* changed from 7, "low average," to 3.5, "good." In half of the cases, pre-training status changed from ranks interpretable as "low average" or "poor" to ranks interpretable as "excellent" or "superior." The number of cases exceeding the median of the standardization population rose from five to thirteen.

5. The music *Ss* markedly reduced their mean percentage of error at each of the five levels of difficulty in the test. The non-music *Ss* greatly reduced their mean percentage of error at increments of 8 ~ and 5 ~, but their mean performance remained relatively stable at the three more difficult increments.

6. The mean score made by the non-music group was higher *after* training than the mean score of the music group had been *before* training.

b. *Wyatt pitch test.* The Wyatt test was administered to each of the sixteen *S*s at least twice before training and at least twice after training. A few of the *S*s had additional tests, making a total of 35 pre-training and 34 post-training testings. As in the case of the Seashore test data, two types of comparison were made, based on (1) average performance in all of the pre-training *vs.* all of the post-training tests taken; (2) performance in the last pre-training test taken *vs.* performance in the last post-training test taken. Only the data derived in the former type of comparison are included here, as the results in the latter types conformed so closely:

1. The mean post-training gain for all sixteen *S*s was 13.60 points, which brought the group about half of the way toward achieving the maximum possible score. The music and non-music *S*s had mean increases of 12.25 and 14.95 points respectively. These mean gains amounted to approximately 56% and 47% of the maximum possible gain for each group.

2. For both groups, the differences between pre-training and post-training means yielded *t* values which were "highly significant" at the 1% level of confidence.

3. All sixteen of the *S*s improved in this test, with three-fourths of the cases scoring 8 or more points higher in their post-training tests.

4. The Wyatt test has different norms for music and non-music *S*s. Relative to their respective populations, each group had an average pre-training rank of 9 ("poor") and each changed to a rank of 5 ("average"). Before training, fifteen of the sixteen *S*s were below the median for their respective populations, but only three failed to equal or exceed the median after training.

5. The music *S*s reduced their mean percentage of error at each of the ten levels of difficulty in the test. The non-music *S*s markedly reduced their mean percentage of error at all levels excepting the 0.5 ~ increment.

6. As in the case of the Seashore test, the mean post-training performance of the music *S*s was superior to that of the non-music *S*s. Here too, however, the test performance of the latter group was better *after* training than the test performance of the former group had been *before* training.

c. *Oscillator tests.* Tests were given both before and after training at three different frequency standards—250 ~, 500 ~ and 1000 ~. Since all of the training in discrimination and practically all of the training in intonation was given at a standard frequency of 500 ~, the tests at this frequency measured improvability at the training level, while the tests given at standard frequencies of 250 ~ and 1000 ~ measured transfer to tones which were not heard or sung during the training period. The results tabulated below show more marked improvement at the 500 ~ level than at the two other levels, but positive and statistically significant transfer to these two other frequencies is also indicated:

	Standard Frequencies		
	250 ~	*500 ~*	*1000 ~*
Percentage of *S*s who improved	100%	100%	100%
Average reduction in error			
Music *S*s	44.8%	73.7%	44.4%
Non-music *S*s	51.0%	75.1%	63.6%
Level of confidence of *t*			
Music *S*s	1%	1%	1%
Non-Music *S*s	2%	1%	1%

d. *Incidence of marked improvement.* In order that comparisons could be made of performance in all five of the tests, results were expressed in relative, as well as absolute, terms. This was accomplished by using as a criterion of improvement the percentage by which *S*s succeeded in reducing the magnitude of their pre-training error and in approaching a perfect performance in each test. Computation of such percentages of improvement for each *S* in each test taken indicated that in the case of fourteen of the sixteen *S*s, marked improvement[52] was made in all, or all but one of the tests taken[53] and that in no instance was there a failure to improve in at least two of the five tests.

---

[52] "Marked" improvement may be arbitrarily defined here as a reduction by 30% or more in the magnitude of pre-training error.

[53] One of the *S*s (U.M.) took only three of the five tests, but improved markedly in all three. All of the other *S*s took all five of the tests.

V. Implications

THE RESULTS in the present experiment indicate that the pitch discrimination of initially pitch deficient adults was significantly improved after intensive training designed to be "remedial," i.e., adjusted to the individual needs of the *S*s. It was also found that training transferred significantly to discrimination of tones at standard frequencies which were respectively one octave lower and one octave higher than the standard tone employed in substantially all of the training.

The following implications are suggested by this study:

1. No practical method has yet been devised, to the knowledge of the writer, for determining with certainty that a given measurement represents an individual's "physiological limit." In the present experiment, even the results in the final post-training retests do not necessarily indicate that the bed-rock limit of "capacity" was measured, for it is possible that a longer or more varied training program might have resulted in still further improvement. Inasmuch as an individual's proficiency in pitch discrimination can only be known through *measurement,* it would appear to be more realistic, for practical purposes, to ignore the concepts of a fixed capacity and of a physiological limit and to regard a threshold, score, rank or any other quantitative designation of proficiency, as indicative of the individual's *ability,* i.e., simply as what he is *able* to do in a given measurement at a given time.

2. If, for practical purposes, the tests of pitch discrimination are regarded as measures of an ability rather than a capacity, the term "cognitive limit" would also seem devoid of function. In fact, the term may actually be misleading. It is not uniformly or satisfactorily defined and it seems to suggest, moreover, that the factors which inhibit maximum performance are concerned only with cognition, while the evidence summarized below indicates otherwise.

3. Although it is true that the tests of pitch discrimination are simplified by isolating pitch as the only variable and by asking *S*s merely to respond "higher" or "lower," simplicity in the content and form of the test does not constitute unequivocal proof that the individual taking the test is responding only at the sensory level even when the test is taken under optimal conditions. Crucial evidence on this point would probably require the application of elaborate neurophysical techniques which have not yet been devised. While we do not as yet understand the exact nature of all of the psychological processes involved in discrimination of pitch or the nature of the changes which take place in *S*s when their pitch discrimination improves, it seems probable that proficiency in a pitch discrimination test may be affected by many factors which are quite remote from simple auditory sensation and which may involve, not just the auditory sensorium, but possibly even the entire organism. This probability is supported by the following lines of evidence:

a. Many of the *S*s reported some form of motor participation and many others gave overt indications of such activity.[54] Although in connection with rhythm, many psychologists have recognized the importance of empathy, "the tendency to feel oneself into the music and act it out" (*26,* p. 144) and Seashore has even stated that "rhythm is never rhythm unless one feels that he himself is acting it out"

[54] *Vide* pp. 41-43.

(*ibid*, p. 142), the possibility of a similar motor attitude in connection with pitch perception has not received adequate attention. It is here suggested that the development of greater proficiency in pitch discrimination may entail the development of an increased readiness to imitate tones inwardly; that the process of discriminating pitch may not be due simply to the action of the cochlea and the auditory nerves, but may depend in part upon the individual's ability to "empathize" tones, i.e., to apprehend tones in terms of mimetic bodily movements.

b. Judging by commentary of many of the *S*s coupled with changes in their performance, it appears that imagery may also have an important influence upon pitch discrimination. It was found that training methods which encouraged the use of more vivid imagery seemed to facilitate pitch discrimination and to lead to more accurate responses.[55]

c. Postural attitudes were also found to be related to performance in pitch discrimination, although the lack of uniformity among *S*s suggests that postural attitudes which are optimal in some instances may be detrimental in others.[56]

d. The conventional "cognitive" factors such as application, motivation, understanding of the test requirements, etc., undoubtedly also influence results in a particular testing. When special training was given, however, improvement was found to occur even when *S*s did not appear to be "cognitively" limited and even after individual retests were carefully administered. In Cameron's experiment, for example, the six *S*s who participated were all trained psychologists for whom such difficulties must have been negligible. Moreover, the tests were preceded by twenty minutes of preliminary practice. Yet improvement in pitch discrimination occurred at the level at which singing practice was given and not at the unpracticed level. Cameron therefore ascribed improvement to the practice in singing.[57] Cognitive difficulties, as defined above, also appear to have been minimal in the present experiment, for all of the *S*s were intelligent adults and half of them were musically trained. Nevertheless, individual retesting prior to training yielded results which were significantly inferior to those which followed the training.

These considerations imply that the act of discriminating pitch, "elemental" as it may seem, may actually involve a complex of psychological functions, including sensory, perceptual, imaginal, motor, intellectual and affective processes.

4. Experimental evidence indicates that improvement in pitch discrimination often resembles the learning of skills involving the formation of new habits, as, for example, the "spurts" which occurred in the attainment of improvement reported in the Wolner study.[58] In general, for the development of tonal orientation, or vivid imagery, or for inculcation of the habit of forming implicit sub-vocal motor sets—in short, for the complete integration of new modes of response which are important in pitch discrimination—a prolonged period of diagnosis and intensive remedial training may be required. No valid conclusions regarding the improvability of pitch discrimination should be drawn from experiments which fail to provide patient, persevering and diversified training.

5. A survey of the experimental literature indicates that many erroneous

[55] *Vide* pp. 41-43.
[56] *Vide* pp. 41-43.
[57] *Vide* pp. 20-21.
[58] *Vide* p. 22.

generalizations have resulted from the failure to define "training" carefully. The absence of improvement upon multiple retesting or following formal instruction in music has mistakenly been regarded as final critical proof that pitch discrimination is not improvable through *any* kind of training.

6. The results of the present experiment indicate that great caution must be exercised in the use of such tests for vocational guidance. While such tests may serve a valid prognostic function for these individuals who make high ratings, in the case of individuals who are relatively deficient, it is believed that there is a real need for a shift of emphasis to *diagnosis* of the causes of the deficiency and to a serious effort to devise remedial training procedures. The primary school would seem to be the most appropriate place for the discovery of latent deficiency in pitch perception and for the application of appropriate remedial methods. Thus, in addition to their usefulness for prognosis when ratings are high, tests of pitch discrimination might serve such important purposes as: (1) locating individuals who are deficient in pitch discrimination, (2) suggesting the area and limits of deficiency and (3) providing a measure of improvement in the case of those individuals who respond to diagnosis and remedial training.

7. The views here expressed are highly consistent with R. H. Seashore's "work methods" hypothesis[59] as well as the Gestalt views expressed by Pratt and Mursell.[60] While divergent from the rather sensationalistic and hereditarian position taken by Seashore in his publications prior to 1940, these views are not in opposition to many of the statements found in his 1940 monograph (*28*). In fact, insofar as Seashore concedes that his tests do not necessarily measure physiological limits and regards them as measures of *abilities* which are subject to improvement through environmental influences, the data and the implications of the present study are confirmative.

8. Essentially this study indicates the importance of a clearer understanding of the processes involved in pitch discrimination and of careful individual diagnosis and remedial training. There is need for considerable further research, e.g., on such problems as the degree of permanence of improvement resulting from training or the relationship between specific work methods and improvement. Studies should also be made of the improvability of other functions such as intensity, time, timbre and rhythm discrimination to determine whether remedial training is of any avail and whether standardized remedial techniques can be developed in these areas.

---

[59] *Vide* pp. 4, 24, 29.
[60] *Vide* pp. 3-4.

BIBLIOGRAPHY

1. CAMERON, E. H. Effects of practice in the discrimination and singing of tones. *Psychol. Monogr.*, 1917, *23*, No. 100, 159-180.
2. CAPURSO, A. A. The effect of an associative technique in teaching pitch and interval discrimination. *J. appl. Psychol.*, 1934, *18*, 811-818.
3. CAPURSO, A. A. Effect of an associative technique in teaching pitch and interval discrimination. Unpublished master's thesis, Univ. Kentucky, 1934.
4. CONNETTE, E. The effect of practice with knowledge of results. *J. educ. Psychol.*, 1941, *32*, 7, 523-532.
5. FARNSWORTH, P. R. The effects of nature and nurture on musicality. *Yearb. nat. Soc. Stud. Educ.*, 1928, *27*, 2, 233-247.
6. FARNSWORTH, P. R. An historical, critical and experimental study of the Seashore-Kwalwasser test battery. *Genet. Psychol. Monogr.*, 1931, *9*, 5, 291-393.
7. FARNSWORTH, P. R. Studies in the psychology of tone and music. *Genet. Psychol. Monogr.*, 1934, *15*, 1-94.
8. FISHER, R. A. *Statistical methods for research workers* (4th Ed.). Edinburgh: Oliver and Boyd, 1932.
9. LARSON, RUTH C. Studies on Seashore's 'Measures of Musical Talent.' *Univ. Ia. Stud. Aims Progr. Res.*, 1930, 2, No. 6.
10. LINDQUIST, E. F. *Statistical analysis in educational research.* Boston: Houghton Mifflin, 1940.
11. MCCARTHY, DOROTHEA. A study of the Seashore measures of musical talent. *J. appl. Psychol.*, 1930, *14*, 437-455.
12. MILES, W. R. Accuracy of the voice in simple pitch singing. *Psychol. Mongr.*, 1914, *16*, No. 69, 13-66.
13. MURSELL, J. L. *Psychology of music.* New York: Norton, 1937.
14. MURSELL, J. L. What about music tests? *Mus. Educators J.*, 1937 (Oct. 24).
15. MURSELL, J. L. The issues of the test discussion. *Mus. Educators J.*, 1938 (May 25).
16. PRATT, C. C. *The meaning of music.* New York: McGraw-Hill, 1931.
17. SEASHORE, C. E. The measurement of pitch discrimination: a preliminary report. *Psychol. Monogr.*, 1910, *13*, No. 53, 21-60.
18. SEASHORE, C. E. & JENNER, E. A. Training the voice by aid of the eye in singing. *J. educ. Psychol.*, 1910, *1*, 311-320.
19. SEASHORE, C. E. The tonoscope. *Psychol. Monogr.*, 1914, *16*, No. 69, 20-69, 1-12.
20. SEASHORE, C. E. Elemental tests in psychology. *J. educ. Psychol.*, 1916, *7*, 81-87.
21. SEASHORE, C. E. & MOUNT, G. H. Correlation of factors in musical talent and training. *Psychol. Monogr.*, 1918, *25*, No. 108, 47-92.
22. SEASHORE, C. E. & TAN, K. The elemental character of sensory discrimination. *Psychol. Monogr.*, 1918, *25*, No. 108, 159-163.
23. SEASHORE, C. E. *Manual of instructions and interpretations for measures of musical talent.* New York: Columbia Graphophone Co., 1919.
24. SEASHORE, C. E. *Psychology of musical talent.* Newark: Silver, Burdette, 1919.
25. SEASHORE, C. E. Two types of attitude toward the evaluation of musical talent. *Mus. Educators J.*, 1937 (Dec. 24).
26. SEASHORE, C. E. *Psychology of music.* New York: McGraw-Hill, 1938.
27. SEASHORE, C. E., LEWIS, D. & SAETVEIT, J. G. *Manual of instructions and interpretations for the Seashore Measures of Musical Talents.* Camden: RCA Mfg. Co., 1939.
28. SEASHORE, C. E., LEWIS, D. & SAETVEIT, J. G. Revision of the Seashore Measures of Musical Talents. *Univ. Ia. Stud. Aims Progr. Res.*, 1940, *65*, New ser., No. 388.
29. SEASHORE, R. H. Improvability of pitch discrimination (Abstract.) *Psychol. Bull.*, 1935, *32*, 546.
30. SEASHORE, R. H. Work methods: an often neglected factor underlying individual differences. *Psychol. Rev.*, 1939, *46*, 2, 123-141.
31. SMITH, F. O. The effect of training in pitch discrimination. *Psychol. Monogr.*, 1914, *16*, No. 69, 67-103.
32. STANTON, HAZEL M. Seashore's 'Measures of Musical Talent.' *Psychol. Monogr.*, 1923, *39*, No. 178, 135-144.
33. STANTON, HAZEL M. & KOERTH, WILHELMINE. Musical capacity measures of adults repeated after musical education. *Univ. Ia. Stud. Aims Progr. Res.*, 1930, 31.
34. STANTON, HAZEL M. & KOERTH, WILHELMINE. Musical capacity measures of children repeated after musical training. *Univ. Ia. Stud. Aims Progr. Res.*, 1933, *42*, New ser. No. 259.
35. STANTON, HAZEL M. Measurement of musical talent. *Univ. Ia. Stud. Psychol. Music,* 1935, 2.
36. VANCE, T. F. Variation in pitch discrimination within the tonal range. *Psychol. Monogr.*, 1914, *16*, No. 69, 115-149.
37. WARREN, H. C. *Dictionary of Psychology.* Boston: Houghton Mifflin, 1934.
38. WHIPPLE, G. M. Studies in pitch discrimination. *Amer. J. Psychol.*, 1903, *14*, 289-309.

39. Wolner, M. *An experiment in individual training of pitch deficient children.* Unpublished master's thesis, Colleges of the City of Detroit, 1932.

40. Wolner, M. & Pyle, W. H. An experiment in individual training of pitch-deficient children. *J. educ. Psychol.*, 1933, *24*, 602-608.

41. Wright, Frances A. The correlation between achievement and capacity in music. *J. educ. Res.*, 1928, *17*, 50-56.

42. Wright, Frances A. The relation of music endowment and achievement tests to teacher selection. Yearb. Mus. Educator's nat. Conf., 1928 (Proc. Mus. Supervisor's nat. Conf., 1928).

43. Wyatt, Ruth F. An historical and critical study of methods of measuring and testing pitch discrimination. Unpublished master's thesis, Northwestern Univ., 1935.

44. Wyatt, Ruth F. A new instrument for measuring pitch discrimination. *Amer. J. Psychol.*, 1936, *48*, 335-341.

45. Wyatt, Ruth F. A note on the use of 'omnibus' training to validate Seashore's 'capacity' hypothesis. *Amer. J. Psychol.*, 1939, *52*, 638-640.

46. Wyatt, Ruth F. The improvability of pitch discrimination. Unpublished doctoral dissertation, Northwestern Univ., 1941.

47. Young, R. W. & Loomis, A. Theory of the chromatic stroboscope. *J. acoust. Soc. Amer.*, 1938, *10*, 112-118.